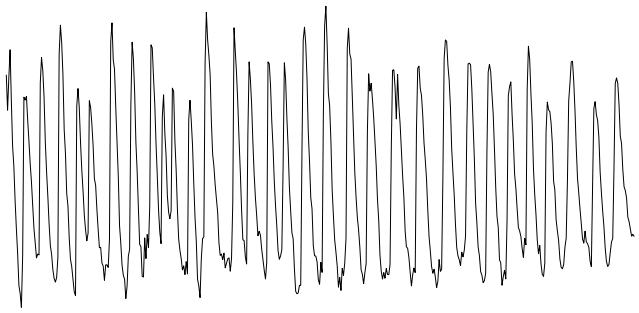
THE ANALYSIS AND MODELLING OF CHAOTIC CALCIUM OSCILLATIONS

SAUL HAZLEDINE



PhD Thesis

University of East Anglia
John Innes Centre
Department of Computational and Systems Biology

July 2010

ABSTRACT

The initiation of the symbiosis between rhizobial bacteria and legume plants is governed by calcium oscillations within plant root hair cells. Using techniques of non-linear time series analysis this work provides evidence that the oscillations are chaotic in nature. An investigation to determine whether the underlying chaotic system can be identified from the oscillations is also described. Finally, a periodic model based on current understanding of the system is presented.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

PUBLICATIONS

Some ideas and figures have appeared previously in the following publications:

Differential and chaotic calcium signatures in the symbiosis signaling pathway of legumes.
Kosuta S., Hazledine S., Sun J., Miwa H., Morris R. J., Downie J. A., Oldroyd G. E. (2008) Proceedings of the National Academy of Sciences USA 105 (28) 9823-8

Nonlinear time series analysis of nodulation factor induced calcium oscillations: evidence for deterministic chaos?
Hazledine S., Sun J., Wysham D., Downie J. A., Oldroyd G. E., Morris R. J. (2009) PLoS One 4 (8) e6637

ACKNOWLEDGMENTS

The author would like to thank his supervisor, Richard Morris, for the help, strategy and many ideas that have contributed to this thesis. The author is also grateful to his co-supervisor Giles Oldroyd for his patience in giving the critical biologists viewpoint on the work being done.

Jongho Sun has provided the majority of the experimental data analysed in this work and has been very responsive in performing and designing additional experiments as well as suggesting conceptual models for the system. The author spent a week under the guidance of Krasimira Tsaneva who provided invaluable assistance and proposed the intial mathematical model given in Chapter 4. Derin Wysham is a major contributor to the work described in Chapter 2 and has advised on the mathematical analyses elsewhere.

Alan Downie has supported the studies through important, useful and open discussions. Emma Granqvist has provided essential references as well as background information and advice on a daily basis. Myriam Charpentier suggested significant improvements to a draft of this thesis, made many helpful comments during the research and proposed the conceptual model that led to the mathematical models in Chapter 4. Pauline Haleux and Emma Granqvist did the majority of the work that led the analysis of buffers that is described in Section 4.4.

Living and laboratory expenses have been generously funded by the BBSRC.

Part I

THESIS

1 INTRODUCTION

1.1 THE LEGUME SYMBIOSIS WITH RHIZOBIAL BACTERIA

1.1.1 *Nitrogen Fixation*

All protein based life requires Nitrogen but few organisms can fix Nitrogen from the air.

The majority of the Nitrogen required by modern agriculture is produced industrially by the Haber-Bosch process which breaks the triple bonds of N_2 at temperatures greater than 600°C and pressures up to 500 atm. This energy intensive process combined with the transport and application of the resulting fertilizer is estimated to consume 1% of the worlds energy supply [51]. Only half of the Nitrogen applied as fertilizer is taken up by crops with the excess either taken back into the atmosphere or leached into water significantly increasing nitrate levels in ecosystems [31].

Biological Nitrogen fixation provides the majority of the Nitrogen required by non-agricultural plants [51]. It does this using sustainable energy sources and does not raise the nitrate levels of the surrounding environment. Fixed Nitrogen is needed to produce the biomass that performs carbon sequestration implying that the natural systems that fix Nitrogen are an important variable in a changing climate [46].

One naturally occurring Nitrogen fixing system exists as an outcome of a symbiosis between rhizobial bacteria with legume plants. This research focuses on the rhizobial/legume symbiosis, in particular, the symbiosis between the model legume *Medicago truncatula* with the bacteria *Sinorhizobium meliloti*.

1.1.2 *Outline of the Legume/Rhizobia Symbiosis*

The majority of plants do not form symbioses with Nitrogen fixing bacteria and the root nodule symbiosis is restricted to four orders within the Eurosid I clade of Angiosperms [62]. One of the best studied symbioses involves legume plants such as *Medicago truncatula* and bacteria collectively known as rhizobia.

The symbiosis of rhizobia and *Medicago truncatula* leads to a developmental change within the plant known as nodulation. Organs, called nodules, are formed when rhizobia have penetrated root cortical cells and are bound in membranes produced by the plant cells. Nitrogen fixation occurs within the controlled environment of the nodules. Nodulation only takes place between specific species of bacteria and specific species of legumes.

Nodulation, and the steps leading up to it, have been discussed in several reviews [30, 104, 72, 91, 92]. It consists of the following events with the approximate sequence:

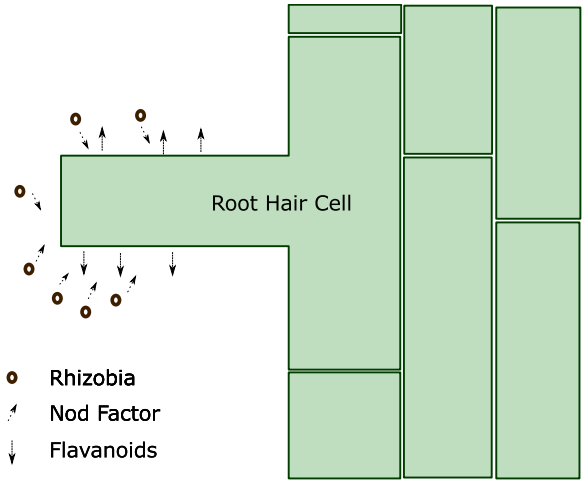


Figure 1: Flavanoid molecules activate secretion of Nod factor.

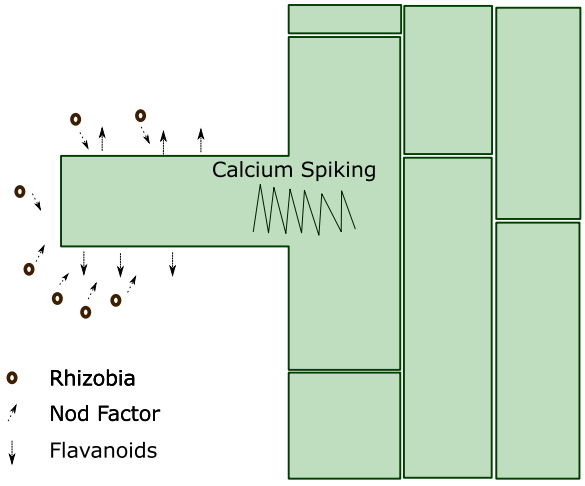


Figure 2: Ca^{2+} spiking occurs in root hair cells within 10 minutes of Nod factor recognition.

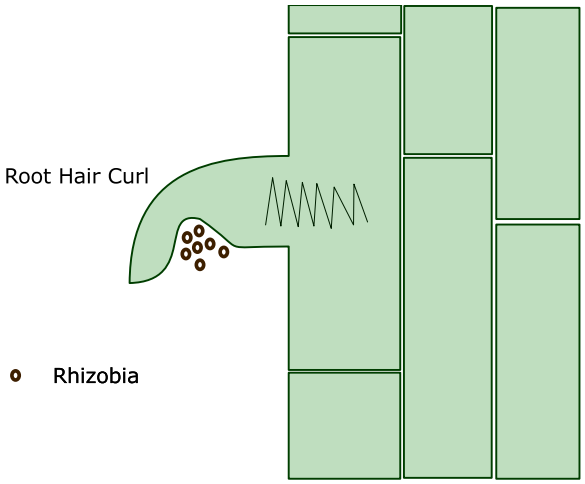


Figure 3: Co-ordinated infection and organogenesis processes follow the sensing of Ca^{2+} oscillations.

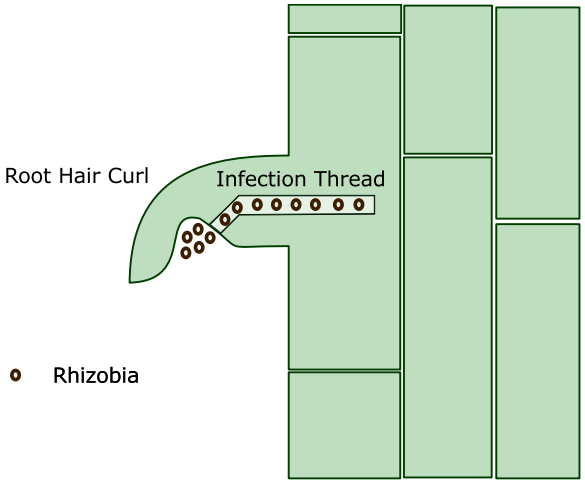


Figure 4: An infection thread forms within the root hair cell.

1. Flavanoid molecules, produced by the host plant root, activate bacterial production and secretion of lipchito-oligosaccharide Nod factors (Figure 1).

2. Nod factor, possibly binding to a receptor, is recognised by the host plant causing intracellular signalling within a root hair cell.

3. Among the earliest events, occurring within 10 minutes, are nuclear localised Ca^{2+} oscillations sometimes known as calcium spiking. The Ca^{2+} oscillations occur within root hair cells (Figure 2).

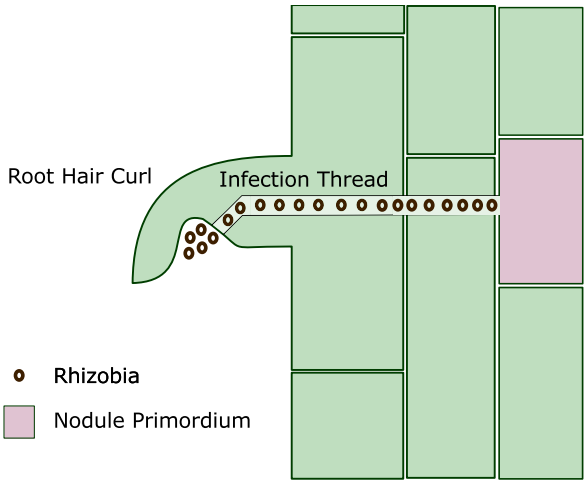


Figure 5: Rhizobia reach the nodule primordium.

4. The sensing of Ca^{2+} oscillations leads to gene expression and the activation of co-ordinated infection and organogenesis processes (Figure 3).
    - Microtubules within the root hair cell begin to disintegrate.
    - Root hair tip growth orientates towards the direction of the Nod factor source.
    - Root hair curling encapsulates rhizobia within a curl.
5. The root cell wall in the vicinity of the bacteria in a curl degrades.
6. An infection thread, constructed from root hair cell wall material, forms within the root hair cell. The infection thread wall is contiguous with the cell wall of the root hair (Figure 4).
7. Rhizobia enter the infection thread.
8. Rhizobia replicate as the infection thread grows and penetrates the epidermal layer.
9. Rhizobia reach the nodule primordium, that has been formed by cortical cells dividing, through infection thread growth (Figure 5). The bacteria are delivered by endocytosis of the infection thread membrane to the nodule primordium where they form symbiosomes. Within symbiosomes they differentiate into Nitrogen-fixing bacteriods.

1.1.3 *The Role of Ca^{2+} Oscillations*

This thesis concentrates on the characterising and modelling of the Ca^{2+} oscillations that occur in root hair cells during the symbiosis between *Medicago truncatula* and *Sinorhizobium meliloti*.

There is genetic evidence to show that Ca^{2+} oscillations and Ca^{2+} signal transduction are essential for the symbiosis to take place. The DMI1 gene (DOESN'T MAKE INFECTIONS) in *M. truncatula*, that encodes a nuclear localised cation channel, is required for both nodulation and Ca^{2+} oscillations [125, 95].

A Ca^{2+} and calmodulin-dependent kinase (CCaMK), encoded by the gene DMI3, is essential for nodulation [71] and can induce nodules when its autoinhibitory domain is removed [35]. The modified DMI3 is also able to induce genes associated with symbiosis in a DMI1 mutant that does not generate Ca^{2+} oscillations [35]. The CCaMK is nuclear localised and a highly Ca^{2+} regulated protein making it a candidate for decoding the Ca^{2+} oscillations and initiating downstream gene transduction.

Symbiosis with Mycorrhizal Fungi

Abuscular mycorrhizal fungi take part in a symbiosis with the majority of land plants obtaining carbon from the plants in return for nutrients, such as phosphate, obtained from their large mycelial networks.

It has been shown that the mycorrhizal symbiosis and the N_2 fixing bacterial symbiosis share common signalling components involved in Ca^{2+} oscillations [62]. DMI1 and DMI3 are required for both the bacterial and mycorrhizae symbioses [2, 71]. Mycorrhizal fungus also induces Ca^{2+} oscillations [65] and induces some of the same genes as N_2 fixing bacteria [53] downstream of the Ca^{2+} spiking. Since plant symbioses with mycorrhizal fungi evolved around 300 million years earlier than symbioses with rhizobia, it is hypothesised that the rhizobial symbiosis recruited existing symbiotic pathways. Some evidence of this can be seen in the SYMRK gene which is required for both types of symbiosis [33] but has more complex structures in plants that form root nodule symbiosis. This extra complexity is required to form the root nodule symbiosis but not the mycorrhizal symbiosis [78].

The Ca^{2+} spiking produced in response to mycorrhizal fungus tends to oscillate at a higher significant frequencies when compared to Nod factor induced Ca^{2+} oscillations [65]. However, because the analysis was done on an unpurified chemical signal, that is not comparable

to using purified Nod factor, it is cannot be determined if defining features of the mycorrhizal Ca^{2+} spiking are due to the intensity of stimulation or differences in the nature of activation. Nethertheless, when taken together with genetic evidence these results show that Ca^{2+} oscillations play a significant role in both symbioses.

1.2 CALCIUM OSCILLATIONS

1.2.1 *The Measurement of Ca^{2+} Oscillations*

Ca^{2+} concentration in a cell is not directly measurable and chemical sensors are required to detect changes in Ca^{2+} concentration. The most popular forms of Ca^{2+} sensors are excited by a laser and change a characteristic of their fluorescence such as wavelength or amplitude when bound to Ca^{2+}. The earliest form of Ca^{2+} sensor used to detect Nod factor induced Ca^{2+} spiking was the dye Fura-2 [24]. Later, Oregon Green dyes [127] or Cameleon proteins [82] were employed.

Upon Ca^{2+} binding Fura-2 changes the wavelength at which it absorbs UV light. The Oregon Green dye increases fluorescence up to $14\times$ in the presence of Ca^{2+} with a wavelength in the visible spectrum [48] which is less perturbing than UV to the cell under study. The Cameleon protein can be added to the genome of plants and can be targeted to different parts of the root hair cell [117]. The Ca^{2+} measurements analysed in this thesis were all obtained by measuring with Oregon Green.

Dyes are normally microinjected into root hair cells resulting in an exceptional addition of Ca^{2+} buffers to the system. In Chapter 4 we investigate the effects of a Ca^{2+} probe on a mathematical model.

Ca^{2+} indicators are excited with a laser during experiments in order to measure fluorescence. Unfortunately this excitation step results in a proportion of the indicators being damaged which is observed as a gradual fluorescence reduction over time known as photobleaching. This can be partially accounted for by using a ratio of fluorescence of Ca^{2+} bound indicator to Ca^{2+} unbound indicator based on the assumption that both modes will photobleach at the same rate. However, photobleaching can still be observed on most Ca^{2+} spiking traces and methods to remove this experimental artifact using detrending are discussed in Chapter 2.

1.2.2 *Observed Behaviour of Ca^{2+} Oscillations*

Oscillations in the level of nuclear and cytosolic Ca^{2+} are observed in root hair cells approximately 10 minutes after Nod factor addition [24]. Using Fura-2 the peak perinuclear Ca^{2+} concentration has been determined as between 433 and 669 nanomolar (nM). This drops to a < 100 nM minimum for basal Ca^{2+} levels in the cytosol.

When oscillations occur, the peak amplitude is more pronounced at the nucleus. Spiking is less pronounced at the root-tip and the pattern of oscillation is also less consistent. This trend can be seen as measurements are taken further away from the nucleus.

After addition of high concentrations of synthesised Nod factor, a Ca^{2+} influx is observed and depolarisation of a root hair cell occurs within 2 minutes. Oscillations do not start until around 10 minutes after the application of Nod factor [24]. However, because lower levels of Nod factor induce Ca^{2+} oscillations without a Ca^{2+} influx [115, 83], it is likely that oscillations occur before the Ca^{2+} influx provided there is a steady increase in the levels of Nod factor produced by the bacteria and perceived by the host plant.

1.2.3 *Frequency, Number of Spikes and Gene Expression*

The induction of early nodulation genes (ENOD11) can be used as a marker for symbiotic development changes in the host plant [53]. ENOD11 genes are not induced throughout the root but only in a zone of developing root hairs. This is in contrast to Ca^{2+} spiking which occurs in root hairs throughout the root with younger hairs at the tip of the root oscillating slower than older hairs further up the root [82].

One possible hypothesis is that the induction of genes depends on the frequency of the Ca^{2+} oscillations. However, roughly halving the frequency of the Ca^{2+} spikes results in slower induction of ENOD11 rather than failed gene induction [82]. This result is consistent with gene induction being connected to the number of Ca^{2+} spikes, possibly due to an integration of the Ca^{2+} concentration during oscillations. It has been estimated that ≈ 36 spikes are required to induce ENOD11.

However, despite the frequency of Ca^{2+} oscillations not being significant for the activation and deactivation of ENOD11, frequency or shape may have a part to play in determining whether a symbiosis is a mycorrhizal or a bacterial one [65].

1.2.4 *Required and Known Components*

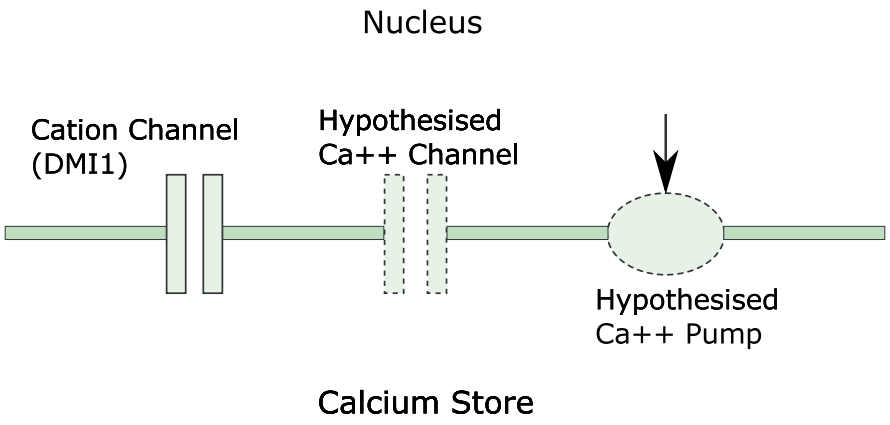


Figure 6: Minimal system required for Ca^{2+} oscillations with known and hypothesised components.

A minimal hypothetical system required for Ca^{2+} oscillations would comprise of a Ca^{2+} store containing a relatively high concentration of Ca^{2+}, a Ca^{2+} channel that releases ions from the store and a Ca^{2+} pump that actively transports ions across the concentration gradient to refill the store. Although promising unpublished data exists, at the time of writing there are no candidates for the roles of Ca^{2+} channel or Ca^{2+} pump.

The modelling of Ca^{2+} oscillations in animal systems is a mature field [113, 28]. However, critical differences exist which suggest that Ca^{2+} oscillations in animals are generated by different mechanisms to those found in plants. In animal systems, Ca^{2+} is released into the cytosol from the endoplasmic reticulum or sarcoplasmic reticulum [8]. In contrast, rhizobia induced Ca^{2+} oscillations in plants show high concentrations of Ca^{2+} around the nucleus [24] and rely on ion channels localised to the nuclear envelope [105, 18]. The majority of models of Ca^{2+} oscillations in animal systems have a behaviour dependent on the complex dynamics of the inositol-1,4,5-trisphosphate receptor (IP3R) [113] for which no homologous proteins have been found in plants [86].

The nuclear envelope is a known Ca^{2+} store that is contiguous with the endoplasmic reticulum and maintains a free Ca^{2+} concentration around $100\times$ that of the nucleus and cytosol [98]. Ca^{2+} release is observed during spiking in the perinuclear region of plant root hair cells, so it is reasonable to suggest that the nuclear envelope is the Ca^{2+} store in this system.

Although no Ca^{2+} channel has yet been identified that takes part in the symbiotic Ca^{2+} oscillations, the putative cation channel encoded by DMI1 is essential for Ca^{2+} spiking to take place. It is unknown how conductive this channel is for Ca^{2+}, K^+ or other cations. In Chapter 4, DMI1 is modelled as a K^+ channel which balances the electric field across the membrane of the nuclear envelope to allow Ca^{2+} to flow.

1.3 OVERVIEW OF THESIS

This thesis consists of three main chapters. Each chapter describes a different analysis of the mechanism, or methods to investigate the mechanism, underlying the Ca^{2+} oscillations that occur during the early stages of symbiosis.

The first Chapter consists of a time series analysis of experimental data. The Ca^{2+} oscillations give positive results in multiple established tests for chaos. This is in contrast to Ca^{2+} oscillations that have been investigated in other systems. A further investigation and discussion of the time series techniques used on other systems is made.

The second Chapter consists of an empirical investigation into whether differential equations, that exhibit chaos, can be recovered from a single chaotic time series generated by the equations. The investigation is unable to demonstrate success with a realistically sized chaotic system. However this Chapter may be of interest to other researchers, as to our knowledge, several techniques are used together and compared for the first time.

At the time of writing, no mathematical models for the symbiotic Ca^{2+} spiking system have been published. The third Chapter describes a periodic model for the Ca^{2+} oscillations. This model is fit to available experimental data and analysed using techniques from nonlinear dynamics.

2

DETECTING CHAOS IN CALCIUM OSCILLATIONS

2.1 OVERVIEW

The chapter describes a nonlinear time series analysis of the Ca^{2+} oscillations that occur in *M. Truncatula* during the early stages of symbiosis with nitrogen fixing bacteria. The results of the nonlinear time series suggest, but do not prove, that the oscillations are chaotic.

The content of this chapter is taken from the paper *Nonlinear Time Series Analysis Of Nodulation Factor Induced Calcium Oscillations: Evidence for Deterministic Chaos?* [42]. The methods section and the supplemental information have been moved nearer to the start of the report for clarity. Some points and explanations have also been expanded.

2.2 ABSTRACT

Legume plants form beneficial symbiotic interactions with nitrogen fixing bacteria (called rhizobia), with the rhizobia being accommodated in unique structures on the roots of the host plant. The legume/rhizobial symbiosis is responsible for a significant proportion of the global biologically available nitrogen (Table 1). The initiation of this symbiosis is governed by a characteristic calcium oscillation within the plant root hair cells and this signal is activated by the rhizobia. Recent analyses on calcium time series data have suggested that stochastic effects have a large role to play in defining the nature of the oscillations. The use of multiple nonlinear time series techniques, however, suggests an alternative interpretation, namely deterministic chaos. We provide an extensive, nonlinear time series analysis on the nature of this calcium oscillation response. We build up evidence through a series of techniques that test for determinism, quantify linear and nonlinear components, and measure the local divergence of the system. Chaos is common in nature and it seems plausible that properties of chaotic dynamics might be exploited by biological systems to control processes within the cell. Systems possessing chaotic control mechanisms are more robust in the sense that the enhanced flexibility allows more rapid response to environmental changes with less energetic costs. The desired behaviour could be most efficiently targeted in this manner,

Table 1: Annual transfer rates of Nitrogen from air to land [37].

Atmospheric	10×10^6 tonnes
Industrial	36×10^6 tonnes
Biological	140×10^6 tonnes

supporting some intriguing speculations about nonlinear mechanisms in biological signaling.

2.3 INTRODUCTION

Calcium oscillations regulate a number of processes in plants, including the establishment of the legume/rhizobial symbiosis. During this interaction, bacteria (called rhizobia) invade the plant roots and are accommodated in membrane bound compartments within plant cells of a specialized organ on the root: the nodule. Within the nodule the bacteria convert atmospheric dinitrogen into ammonia, a form of nitrogen readily available to the plant. The availability of nitrogen is one of the most limiting factors for plant growth and fixed nitrogen from the legume/rhizobial symbiosis provides an essential nitrogen source for agriculture and natural ecosystems.

The establishment of the legume/rhizobial symbiosis involves a molecular communication between the plant and the bacteria, with bacterially-derived Nod (nodulation) factor acting as a central signal to the plant. Perception of Nod factor by legumes activates most of the developmental processes associated with the formation of a nodule. The Nod factor signal transduction pathway of legumes has been well characterized and involves calcium oscillations, termed calcium spiking. An example of calcium spiking is given in Figure 7. Receptor-like kinases are involved in the perception of Nod factor and this leads to induction of calcium spiking via cation channels, that appear to regulate potassium movement and components of the nuclear-pore complex [91]. This signal transduction pathway has also been shown to function in the establishment of a second symbiotic interaction: the mycorrhizal symbiosis. This interaction involves the colonization of the plant root by mycorrhizal fungi that aid the plant in the uptake of nutrients from the soil. Mycorrhizal fungi have been shown to activate calcium oscillations, but with a different structure to Nod factor induced calcium spiking [65]. This suggests that the symbiosis signaling pathway can be differentially activated by both rhizobia and mycorrhizal fungi.

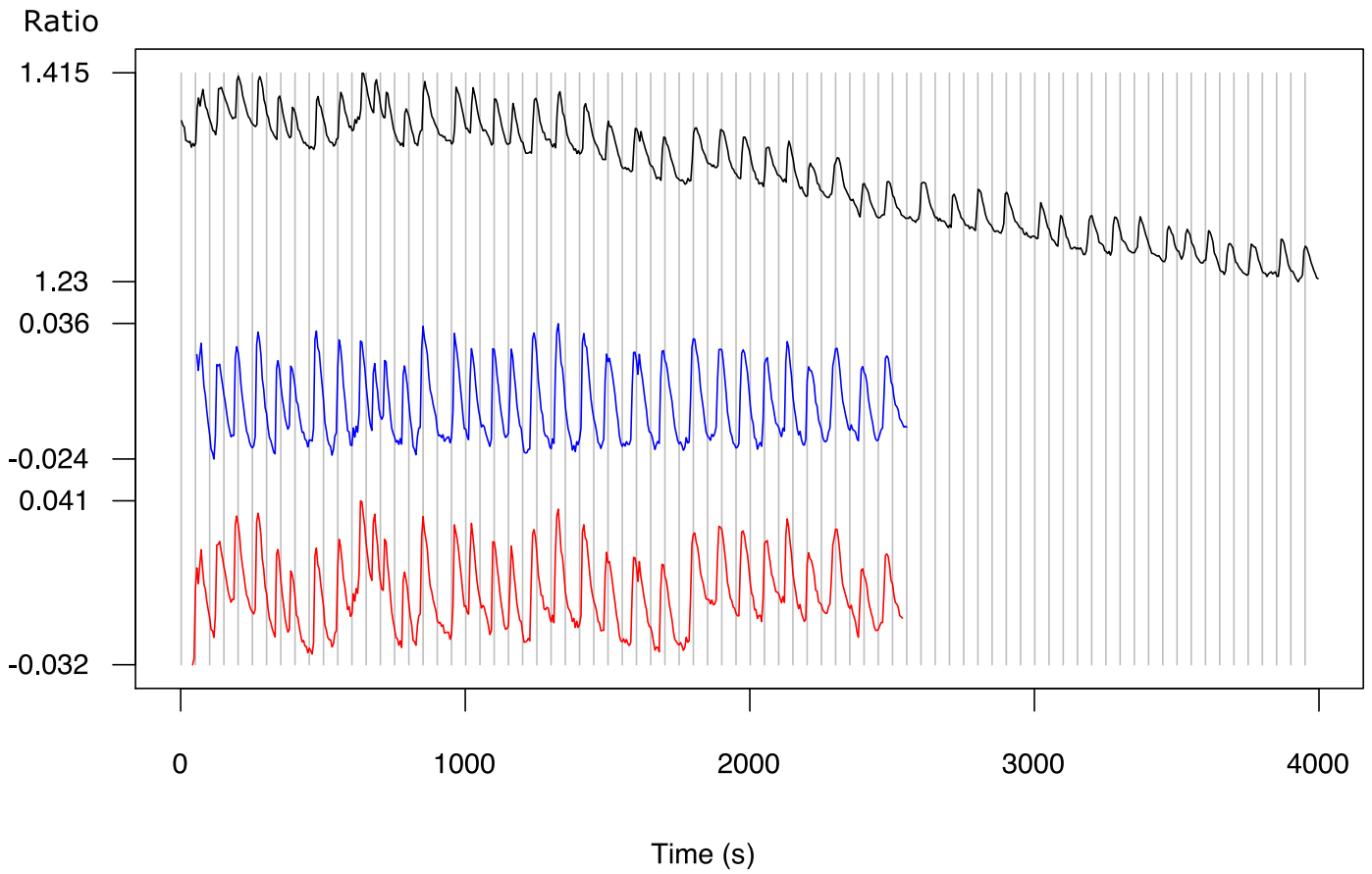


Figure 7: Time series Nod1 given as an example of a raw Nod Factor induced Ca^{2+} spiking trace and after detrending using a moving average (blue) and Empirical Mode Decomposition (red). The Y axis is a fluorescence ratio between Ca^{2+} sensitive and Ca^{2+} insensitive dyes. The X axis is time in seconds.

The nature of biological systems and the challenges inherent in experimentation often result in seemingly erratic time-series behaviour with little apparent structure. Despite advances in signal processing methodology, the extraction of information from such data remains a challenge. Erratic behaviour is often thought to be the consequence of noise or stochastic effects, but apparent randomness can also be generated by a deterministic system operating in the chaotic regime. A universally accepted definition of chaos is still outstanding, however, a number of key features are held in common: A chaotic system is deterministic, nonlinear, and highly sensitive to the initial conditions. The exponential divergence of nearby trajectories implies that the predictability is limited to short time scales. Long term forecasts become impossible despite the underlying deterministic nature. Unpredictable systems are frequently handled with the methods of probability theory and termed stochastic.

Sophisticated techniques exist for distinguishing between linear, nonlinear, deterministic, stochastic and chaotic systems. However, disentangling experimental noise, stochastic effects, and underlying deterministic laws is non-trivial and the initial data derived from biological processes are not often of sufficient quality to allow such analyses. Experimental investigations into calcium (Ca^{2+}) oscillations have frequently been accompanied by mathematical modeling and a wide range of models exist (see [113] for an excellent review of this topic). Questions, however, remain as the mechanisms responsible for the Ca^{2+} signal en- and decoding likely vary between organisms and are not fully understood.

For example, intracellular Ca^{2+} oscillations and Ca^{2+} spikes have been modeled with chaotic systems [13, 67, 39], although stochastic descriptions have been proposed for some of the ion channels involved [29]. Initial chaotic models were inspired by the bursting behaviour observed in experiments on hepatocytes [20, 17, 36]. However, a later theoretical study has shown that an example Ca^{2+} oscillatory system can only be modeled deterministically at physiological Ca^{2+} concentrations when bursting is not taking place [68].

A recent study on Ca^{2+} oscillation data from hepatocytes, which included bursting, led to the conclusion that calcium oscillations were predominately stochastic in nature [97]. Time series data from four cell types in mice and humans was used to show a rapidly falling autocorrelation between Ca^{2+} spike intervals [119]. This was interpreted as evidence that Ca^{2+} spikes are initiated stochastically.

Further analysis revealed that the statistics of the interspike intervals are in agreement with a stochastic model.

In plants, moreover, little is known about the secondary messengers or calcium channels that may direct Nod factor induced calcium spiking [92], also it is apparent that there are major differences in the proteins that activate or perceive well-characterized animal secondary messengers such as inositol trisphosphate and cADPR [86]. Given these unknowns and differences, we are reluctant to bias our analysis towards the models and conclusions drawn from animal systems. Instead, a more appropriate approach to understand Nod factor signaling is to analyse the experimentally obtained calcium oscillations using methods from nonlinear time series analysis. Using a series of techniques, we demonstrate that Nod factor induced Ca^{2+} oscillations generated within the legume M. truncatula are deterministic, nonlinear and show an exponential divergence that is typical of chaotic systems. This observation suggested an alternative explanation to a stochastic interpretation and prompted us to validate our methodology using negative and positive controls. We generated time series using the chaotic Lorenz system of differential equations and the chaotic Haberichter model of Ca^{2+} oscillations. These models were tested alongside our experimental data. We find that while both these positive control data sets would be classified as chaotic using many classical methods, they would be categorized as stochastic using the methods employed in recently published time series analyses of Ca^{2+} oscillations. Whereas stochastic modeling is often an effective approach, the extrapolation from a modeling convenience to the nature of observed phenomena is not without risk and interesting phenomena may be overseen and/or ascribed to random effects. We therefore take a number of precautions to present as thorough an analysis as possible of the experimental Ca^{2+} oscillations.

2.4 MATERIALS AND METHODS

2.4.1 *Time Series and Controls*

We analysed time series data obtained from root hair cells of *M. truncatula* treated with 10-9M Nod factor from *S. meliloti*. The nature of the Ca^{2+} oscillations is comparable whether the plant is treated directly with Nod factor or with *S. meliloti* [126], but for ease of experimentation in this study we have chosen to use isolated Nod factor. The changes in Ca^{2+} levels were measured using the ratio of fluorescence

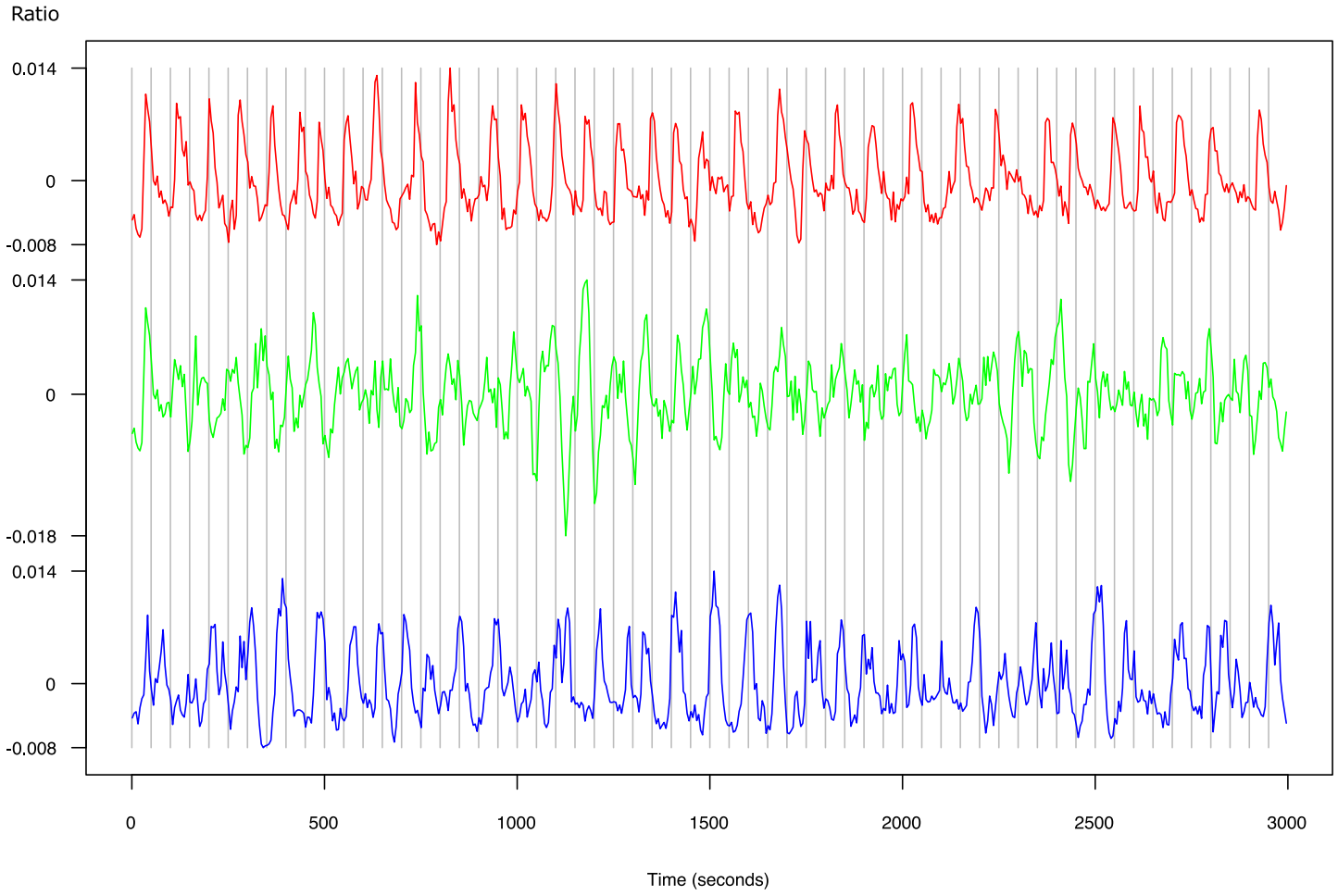


Figure 8: Examples of an experimental time series after moving average detrending (red), an AR model fitted to the experimental data (green) and a surrogate time series (blue). Y axis is a fluorescence ratio between Ca^{2+} sensitive and Ca^{2+} insensitive dyes.

from two dyes: Oregon Green that responds to calcium levels with changes in its fluorescence and Texas Red that is not responsive to calcium and provides a control for fluorescence changes unrelated to calcium. These dyes were micro-injected into root hair cells and fluorescence measured as described in [121]. The intensity of the fluorescence was measured in individual cells at five second intervals for a period of at least 60 minutes. Examples of an unprocessed time series and detrended time series are given in Figure 7.

Experimental time series from 9 cells, were analysed. After detrending by two methods, splitting some time series according to stationarity tests and removing one EMD detrended time series due to nonstationarity, we were left with a total of 21 Ca^{2+} spiking time series labelled 'Nod1 MA', 'Nod1 EMD', ..., 'Nod11 EMD' (described in Table 2 and shown in Appendix 1).

The comparison of the autocorrelation of interspike intervals used two time series obtained from chaotic mathematical models as positive controls. One of the positive controls was generated by a model of Ca^{2+} spiking developed by Haberichter et al [39] and the other by the well known chaotic Lorenz system [76]. Tests for determinism used a time series generated with random numbers obtained from http://www.randomnumbers.info/ as a negative control. As negative controls for nonlinearity we produced two time series, an instance of an autoregressive (AR) model, and a surrogate [112], from each experimental time series (Figure 8). To see the effects of a time series analysis on the type of system suggested by [119], a simple nonlinear model with random interspike intervals was tested.

Autoregressive model

The first type of negative control we used is the result of an autoregressive (AR) model of the form:

$$x_t = a_1 x_{t-1} + \cdots + a_p x_{t-p} + \epsilon \tag{2.1}$$

where x_t is a value of the time series at time t and the model consists of p terms. The value ϵ is a Gaussian term with a mean of zero and a variance σ. An AR model is fitted to the experimental time series by calculating the coefficients $a_{1..p}$ which give the best fit to the experimental time series using the Yule Walker equations [16]. The order p is specific to each time series and is chosen with the Akaike Information Criterion [1]. A new time series is generated with ϵ equal to the prediction variance of the fitted model.

Surrogate time series

Another type of negative control is a time series with the same power spectrum and amplitude distribution as the original time series [112]. A matching power spectrum ensures the negative control has the same linear statistics as the experimental time series. A matching amplitude distribution ensures the negative control has been subject to the same, constant, possibly nonlinear, measurement function as the original time series. Additionally, the phases of the Fourier transform of the original time series are randomised so that the negative control, or surrogate, does not have a deterministic nonlinear component.

The resulting surrogate is a linear Gaussian process with strong similarity to the original time series. Surrogate time series are also used in their more traditional role as described in a later section of this analysis to perform Monte Carlo testing for nonlinearity.

Random Interspike Intervals

The following model was used to generate a synthetic time series as a negative control. The model describes time series measurements x_t at time t. To generate the time series the model has a state $S \in \{\text{spike}, \text{release}\}$, the time since the last spike, τ, total spikes n, and a set of interspike intervals that follow a normal distribution, $\{\alpha\}_{i=1}^{n-1} \sim N(\mu, \sigma)$:

$$x_t = x_{t-1} + k_1 \quad \text{when } S = \text{spike} \tag{2.2}$$

$$x_t = x_{t-1} + k_2\tau \quad \text{when } S = \text{release}. \tag{2.3}$$

The model produces a linear spike followed by an exponential decay. The state changes from spike to release when x_t exceeds a threshold value. Stochasticity is introduced by changing the state from release to spike when $\tau = \alpha_i$. The shape of the spikes are controlled by the constants k_1 and k_2 with k_2 having a negative sign.

2.4.2 *Detrending*

Motion of the cell cytoplasm, known as cytoplasmic streaming, causes relocalisation of the fluorescent dye and this coupled to photobleaching causes noticeable Ca^{2+} independent changes in the overall fluorescence. The ratio of the Ca^{2+} responsive dye, Oregon green, to the non responsive dye, Texas red, reduces some of these non-specific fluctuations, but does not remove all Ca^{2+} independent changes in fluorescence. To remove these effects a moving average was taken and

the result subtracted from the time series [16]. The number of points in the moving average was particular to each trace and was set to either 19 or 25 points. Changing the number of points gave control over the type of features to be removed. The moving average is a linear method and can obscure non-linearities within the signal. A further danger arises in that human judgment of how many points to include in the moving average may affect the final results. Because of this potential for bias, the Nod factor induced spiking time series were also separately detrended by Empirical Mode Decomposition (EMD) [45]. This method of detrending deals more formally with the nonlinearity of the time series and does not distort the shape of the Ca^{2+} spikes [130], however like most automatic methods it is unable to apply heuristic information and could fail to remove experimental idiosyncrasies.

2.4.3 *Time Delay Embedding*

Phase Space

Many nonlinear techniques operate on a phase space representation of a time series. Phase space is a higher dimensional representation than the one-dimensional measurements that make up a time series. Each point in phase space represents a state of the system being measured. A trajectory through phase space represents the evolution of the system through time. In some dynamical systems, the trajectories are attracted to a structure in phase space known as the attractor. The attractors of chaotic dynamical systems are termed 'strange', for example because the attractors have a dimension which is given by a fraction rather than an integer.

To see why a phase space representation is needed, we consider a mathematical model of Ca^{2+} spiking. One of the simplest mathematical models for Ca^{2+} spiking is the one pool model [21]. In this model, Ca^{2+} oscillations are specified using two equations, one for cytostolic Ca^{2+} concentration ($[Ca^{2+}]_{cyt}$) and one for Ca^{2+} concentration in the ER ($[Ca^{2+}]_{er}$). A single state in this model can be specified using two measurements, $[Ca^{2+}]_{cyt}$ and $[Ca^{2+}]_{er}$. As is common in experimental results, there may only be one recorded observable, for example a time series for $[Ca^{2+}]_{cyt}$. However, the properties of the true dynamics for the one pool model can only be revealed when the time series is embedded in at least 2 dimensions.

The technique of time delay embedding provides a method to construct a phase space representation from a single time series. It

does this by sliding a window, containing more than one measurement, down the time series. Each instance of the window then represents a point in phase space.

False Neighbours

To obtain an accurate representation of phase space from incomplete data, the width of the window used in time delay embedding, is taken to be more than twice the box counting dimension of the attractor [110]. However, the dimensions of the attractor are not known. In order to choose a window width, or more formally an embedding dimension, the time series were embedded with different dimensions and statistics run on each embedding to decide on the most suitable dimension.

The false nearest neighbours algorithm [44] was used to suggest the embedding dimension. The algorithm is based on the observation that two trajectories in phase space, that are close to each other, will remain close to each other one time step into the future. If two points in phase space appear close to each other but then move far apart after moving forward in time, they are known as false neighbours. False neighbours can be caused by noise or by an incorrect embedding dimension. The percentage of false nearest neighbours was graphed for embedding dimension, m, where $2 \leqslant m \leqslant 10$. The embedding dimension $m = 6$ was chosen after reviewing all the traces for a dip in the percentage of false nearest neighbours.

Delay Time

A delay time of fifteen seconds for the embedding was suggested by three different methods: mutual information [56], a drop in autocorrelation to $\left(1 - \frac{1}{e}\right)$ [107] and considering time window length [66].

2.4.4 *Stationarity*

Nonlinear time series analysis treats data as if it has come from a dynamical system consisting of variables, such as cytosolic Ca^{2+} concentration, and parameters such as rate constants. Variables have different values over time whereas parameters are assumed to be fixed. If a parameter changes over the course of a time series, it affects the results of the analysis and the time series can be considered as being

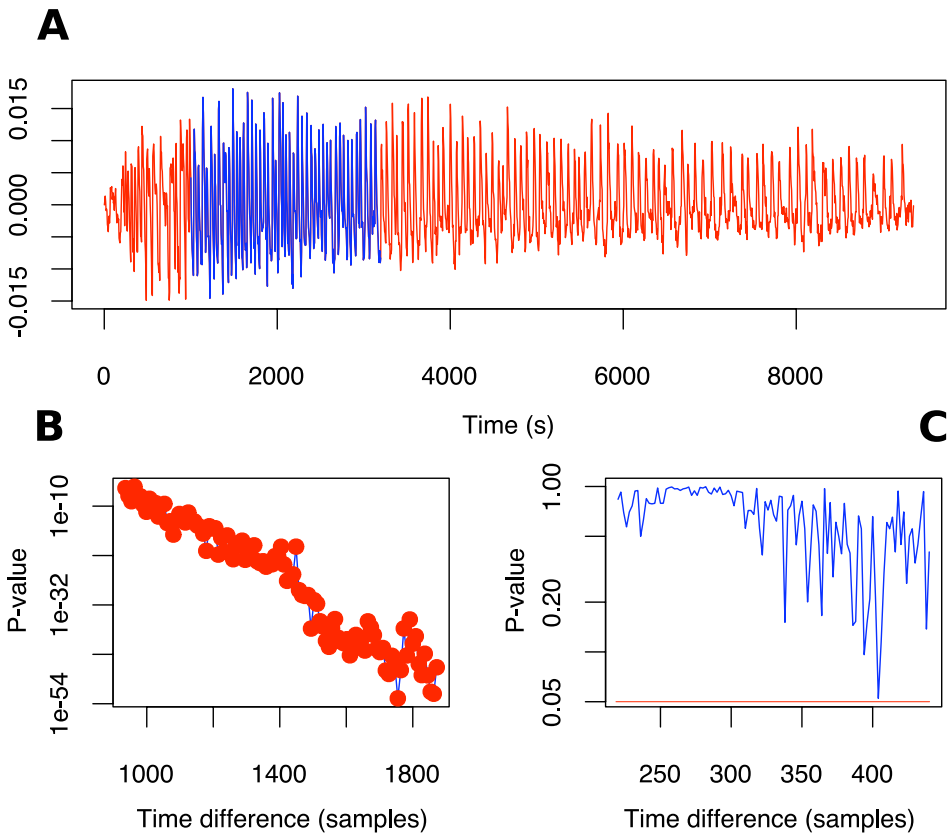


Figure 9: Example of a time series truncated for stationarity. a) The original time series is given in red and the time series, after trunctation, is given in blue. b) The cluster of p-values (y axis) that indicate nonstationarity when the entire series is analysed. c) The results of the stationarity test after truncation. The red line marks the p-value 0.05 which is used as a cutoff for clusters of non-stationary p-values. When a p-value is calculated that is < 0.05 it is marked with a red dot.

non-stationary. A test for non-stationarity [60] was run on each time series to see if any parameter changes could be detected.

The algorithm analyses nearest neighbours. A value D is calculated for nearest neighbours in phase space:

$$D = t_{xnn} - t_x , \tag{2.4}$$

where t_x is the time point that a point in phase space, x, was collected and t_{xnn} is the time that the nearest neighbour to x in phase space was collected. Strands of trajectories are collated where strands are sets of pairs of points which have the same D. If 2 strands share the same point then one of the strands is randomly deleted. The strands are then analysed. The observed distribution of D is compared to stationary system and a *p*-value is calculated that gives an estimate of the probability that the time series is stationary.

We ran the test along various lengths of the time series. Whenever there was evidence of a parameter change, given by a cluster of *p*-values below 0.05 along a section of the time series, the series was cropped before the section showing the parameter change. An example stationary test is shown in Figure 9 where a section of a time series is extracted based on the results of the test.

2.4.5 *Detecting Determinism*

Recurrence Quantification Analysis

A feature of deterministic systems is that they show approximately repeating behaviour. Two nearby trajectories in phase space will remain close in a deterministic system, even if this is only for a short time due to exponential divergence as described below. When working with a time series, a lack of determinism will manifest itself as a rapid reduction in repeating patterns.

Recurrence plots are able to clearly show patterns in data that are missed by viewing the time series alone [22]. These plots are done on a time delay embedded phase space. Where two parts of a time series have neighbouring points in phase space, this is marked by a black dot, or recurrence point, on a recurrence plot. The dots have an effect of marking sections in the time series, that have similar shapes, with diagonal lines.

We used recurrence plots to find approximately repeating patterns in the fluorescence traces. However, interpreting a recurrence plot requires a qualitative step, preferably by somebody who has experience of finding determinism in this way. In order for recurrence plots to be

useful to a wider range of researchers, a quantitative test is needed. We applied a collection of statistical tests, that are run on a recurrence plot, in order to class time series as deterministic [3].

The three tests for determinism use statistics on the number of recurrence points, and the diagonal lines that they form, comparing them to the null hypothesis that a random process generated the time series. As the number of diagonal lines increases, the diagonal lines get longer or a bigger proportion of recurrence points form diagonal lines, the probability that a deterministic system generated the time series increases. The first test operates on the average number of points per diagonal line. The second test looks at the proportion of loose recurrence points that don't form diagonal lines. The third test takes the ratio between the total number of recurrent points and the total number of diagonal lines.

Kaplan-Glass Test

A more established test for determinism [58, 57] analyses the geometry of the time series in phase space. Deterministic trajectories in phase space will have similar orientations to each other. The phase space is separated into an m dimensional grid. Every time a trajectory moves through a box, j, in the grid, the vector from its box entry point to its box exit point, $\hat{v}_{k,j}$, is constructed to have unit length. The average of all the passes through each box $\vec{V}_j$ are calculated:

$$\vec{V}_j = \sum_k \hat{v}_{k,j}/n_j \ , \tag{2.5}$$

where n_j is the number of passes through box j. If trajectories are moving in similar directions, $\vec{V}_j$ will preserve the unit length of each $\hat{v}_{k,j}$ added to it. For perfectly aligned trajectories, the length of $\vec{V}_j$, denoted $|\vec{V}|$, will tend towards 1.

The average values of $|\vec{V}|$ for a given number of passes through a box, n, generated by a random walk through phase space, is available analytically. This value, $\bar{R}_n$ has been shown, numerically, to closely match the $|\vec{V}|$ values for a random signal. To test for determinism, plots of the mean of $|\vec{V}|$, denoted $\bar{L}_n$, against number of passes, n, are compared to a plot of $\bar{R}_n$ against n. Additionally, a single value known as the determinism factor, $\bar{\Lambda}$, can be calculated as a weighted mean of $|\vec{V}|$ and $\bar{R}_n$ as follows:

$$\bar{\Lambda} \ = \ \frac{1}{\sum_j n_j} \sum_j n_j \frac{(\bar{L}_{n_j})^2 - (\bar{R}_{n_j})^2}{1 - (\bar{R}_{n_j})^2} \ . \tag{2.6}$$

For a deterministic system $\bar{\Lambda} \to 1$ and for a random walk through phase space $\bar{\Lambda} \to 0$.

2.4.6 *Testing for Nonlinearity*

Not all nonlinear time series are chaotic. However, in order for a time series to be chaotic it must be nonlinear. Therefore showing nonlinearity is a prerequisite for evidence of chaos. A test for nonlinearity is also a test for nonlinear determinism and analysing nonlinearity reduces the need to run an explicit test for determinism.

Nonlinear Predictions on Surrogate Time Series

After assessing many different tests for nonlinearity, we found the most robust method in the presence of noise to be a test that utilises a non-linear predictor [56]. A non-linear predictor should make better forcasts using a nonlinear time series than a linear time series with random nonlinear effects. We used a conventional combination of a locally constant predictor and numerically generated surrogates. Some nonlinear features of the original series, that do not appear to be due to a nonlinear measurement function, become randomised in the surrogates [112]. The randomisation of nonlinear components transforms the surrogates into linear time series with stochastic effects. For a nonlinear time series, it is expected that the predictor will perform better on the original series than the surrogates.

The nonlinear forecasting algorithm, based on the locally constant predictor, operates on an embedded phase space where it uses points in the phase space as a database to make predictions. A locally constant predictor run over an experimental time series, usually required over 400 datapoints to show an increase in the relative performance compared to a group of surrogates. We speculate that this is because the locally constant predictor requires a threshold number of points in phase space in order to make useful predictions.

In this text, the locally constant predictor is known as the 'zeroth' predictor and using the predictor to test for nonlinearity with a set of surrogate time series is known as the 'zeroth surrogates test'.

2.4.7 *Noise Titration*

Nonlinear tests can be used to test for chaos in conjunction with a technique known as a noise titration [101]. Additive Gaussian noise is

applied to the time series under test to find the point where nonlinearity cannot be detected. It has been demonstrated that the amount of added noise has a relation to the Lyapunov exponent of the system producing the time series. We use the surrogates test for nonlinearity when performing noise titrations.

2.4.8 *Exponential Divergence*

Chaotic systems display a sensitivity to small changes in initial conditions or previous states. A small perturbation ϵ will cause the system to move through a trajectory in phase space that diverges rapidly from the unperturbed trajectory. The sensitivity to perturbation or initial error is quantified as an exponential increasing distance, e^{λ}, between the new trajectory and the original one in phase space. The value λ is known as the maximal Lyapunov exponent. If $\lambda > 0$ the trajectories are diverging, if $\lambda = 0$ the trajectories are parallel and if $\lambda < 0$ the trajectories are converging. The sensitivity of chaotic systems is characterised by maximal Lyapunov exponent being positive, $\lambda > 0$.

Direct Method

It is possible to measure the maximal Lyapunov exponent, λ, from an existing time series even when there is no option of perturbing the system. Some points are chosen from different parts of the time series that are close to each other in phase space. Each point in the neighbourhood will be on different trajectories. The natural log of the average distance between the points is measured then plotted with the natural log of the average distance of the trajectories at various times in the future. Exponentially increasing distance over time, or divergence, appears as a straight line on this plot. The linear trend on the plot enables it to be analysed using linear techniques such as correlation, to indicate if the divergence is indeed exponential, and linear regression, to get a maximal value for λ from the time series.

Measuring the maximal Lyapunov exponent directly from a time series is complicated by the presence of noise. With stochastic effects, $\lambda \to \infty$ and this can distort the linear trend when analysing a system that has exponential divergence. An even greater danger is that non-chaotic systems can appear to have $\lambda > 0$.

The maximal Lyapunov exponent was calculated using a method proposed by Rosenstein [107] suited to short time series of lengths in the order of 1000 points.

Indirect Method

A complementary technique for calculating Lyapunov exponents is the indirect method. In this method, a nonlinear model is fitted to the time series under test and then analysis is performed on the model. However, the accuracy of the maximal Lyapunov exponent will depend on how well the model fits the data. In order to get a good fit to the data many models are fitted. A criterion is used to score how well each model fits the data considering how many parameters were required to produce the fit. In our tests we used the Bayesian Information Criterion (BIC).

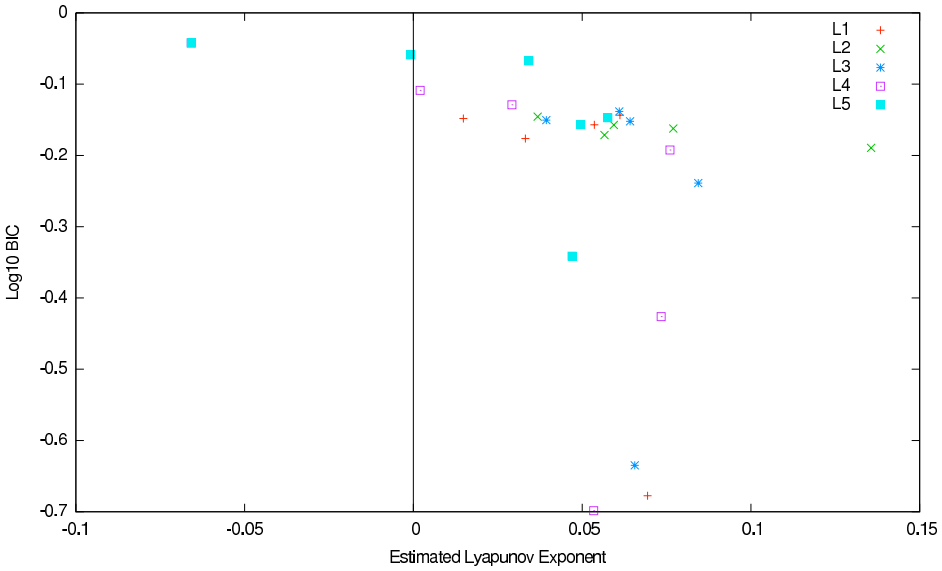
The nonlinear models were fitted using a neural network and have the form [25]:

$$x_n = f(x_{n-l}, x_{n-2l}, \ldots, x_{n-dl}) + \epsilon \,. \tag{2.7}$$

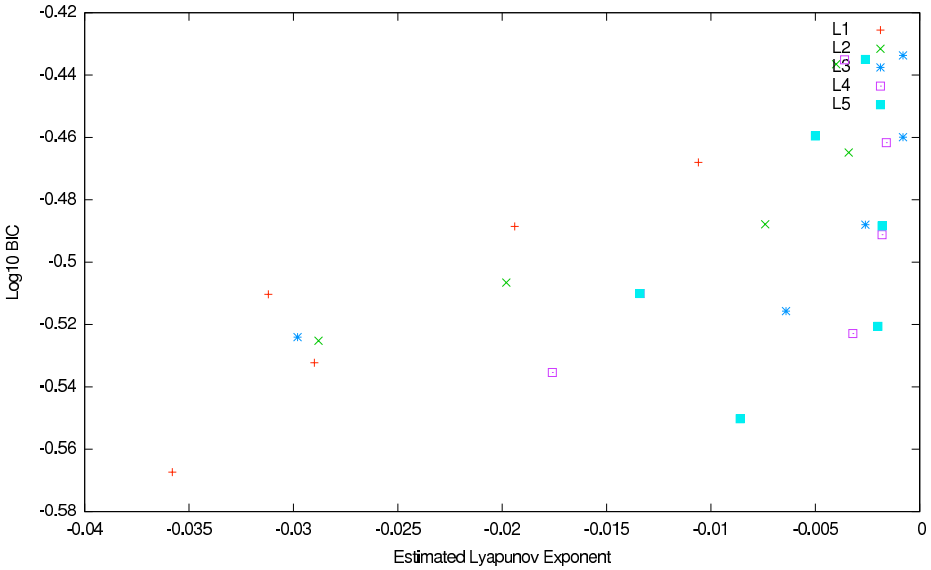
In the model above, l is the delay or lag, d is the dimension and ϵ is a noise term. The models were fitted for various values of l and d. The maximal Lyapunov exponent, λ, was calculated for each model and the signs of the λ calculated for the best fitting models were plotted and considered. An example for an experimental time series and negative control are given in Figure 10 where the experimental time series shows positive exponents as model fitting improves and the negative control shows a weaker fit due to stochasticity and with a majority of λ coming out negative.

2.5 RESULTS

In the following we describe the results of a number of nonlinear time series analyses. In order to check whether a stream of data has arisen from a chaotic system, a number of tests must be carried out. Definitive answers are rare unless the system of underlying equations or map is known. Plotting system observables as a function of themselves at an earlier time gives rise to the return map, which often appears as a simple curve for deterministic systems. The shape of such a curve strongly indicates the classification of the dynamics. This technique is in fact a form of state space reconstruction, in which typical deterministic trajectories should establish themselves upon a low-dimensional attractor. A further test is for exponential divergence and the calculation of Lyapunov exponents, which if positive indicates chaos. These tests are sensitive to noise, which is always present, especially in biological data, and hence rarely provide definitive answers. One of the key steps for such analyses is proper embedding



(a) Experimental Exponents



(b) AR Model Exponents

Figure 10: Examples of indirect Lyapunov exponents calculated for a) an experimental time series b) an AR model used as a negative control. The 'L' values indicate the delay time or lag that was used for particular models and the points are only plotted for the best fitting model for each dimension and lag.

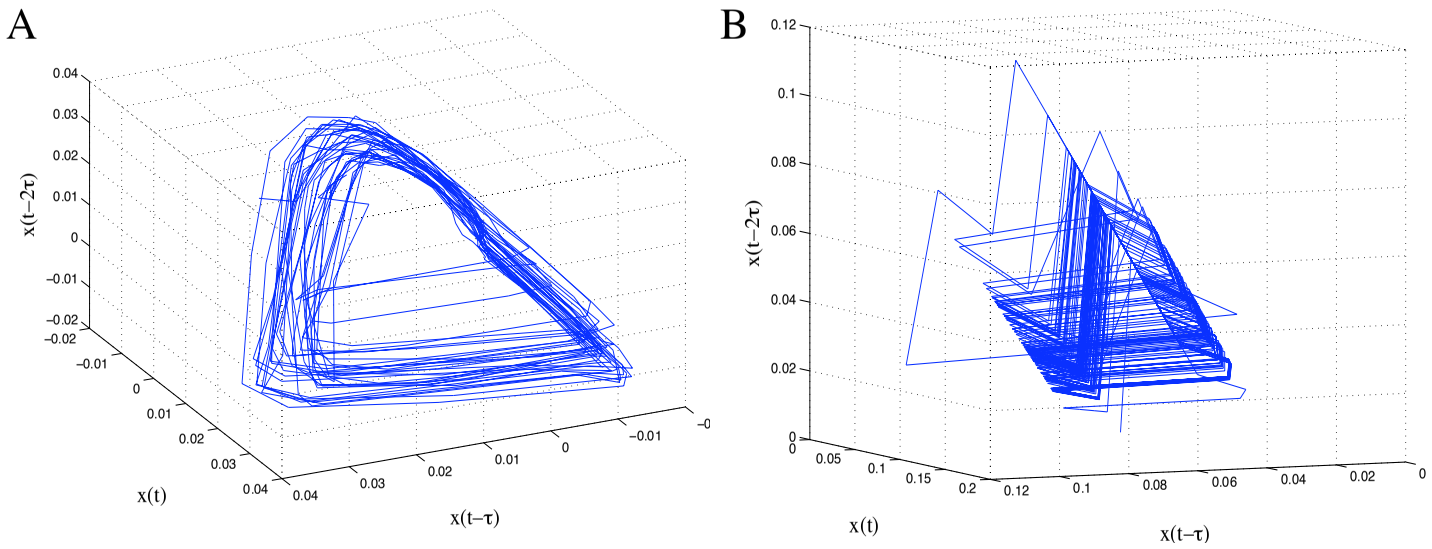


Figure 11: Three dimensional embeddings of a time series of fluorescence ratios (Nod1) and the stochastic spiking model data for comparison. A) The experimental series is clearly noisy, prohibiting accurate dimensionality determination, but it unfolds well to the eye in three dimensions. B) The data from the stochastic spiking model, however, appears to cross itself in many places and coalesces, violating the uniqueness property of Ordinary Differential Equations.

and the determination of attractor dimensionality. Current approaches for these steps work well for data within the order of 2% noise but perform unreliably for noisy data sets. Thus, we are limited in the application of such methods and as a result could not determine the dimensionality reliably, and the return map computations did not produce convincing results. However, as can be seen in Figure 11, the attractor does appear to unfold well in three dimensions. Additionally, a number of tests did provide useful results with a good confidence level. The following sections describe the application of a number of different tests, which taken together certainly do not prove that the Ca^{2+} oscillations are chaotic but do provide evidence that the system could be chaotic.

The results are described in two sections. The first section provides accumulated evidence for chaotic behaviour in the Ca^{2+} time series in M. truncatula. In the second section, additional tests allow a comparison to previous time series analyses that were performed on Ca^{2+} oscillations in animals.

2.5.1 *Evidence of Chaos?*

We analysed the Ca^{2+} oscillations by following the procedure illustrated in the flowchart of Figure 12. The full time series are used and not just interspike times. The time series of Ca^{2+} concentration were first detrended using two different techniques, Empirical Mode Decomposition (EMD) and a moving average, examples of which are shown in Figure 7. Using EMD does not distort the shape of the Ca^{2+} spikes and does not remove low frequency components of the experimental signal. However, because the low frequency components of the signal may not be significant, as an alternative to EMD we also detrended the data using a moving average.

Each detrended Ca^{2+} spiking time series was tested for nonlinearity using a nonlinear predictor and linear surrogates. If nonlinearity was detected, a noise titration was used to test for chaos and the Lyapunov exponent was calculated using a direct method. The direct method calculates the maximal Lyapunov exponent and inspection of the resulting divergence data can help one to discern if the divergence of the system is due to chaotic or stochastic effects. An indirect method was also used where multiple nonlinear models were fitted to the experimental data and a maximal exponent calculated for each model. The indirect method gave a selection of Lyapunov exponents and if a clear majority of well fitting models had positive exponents then

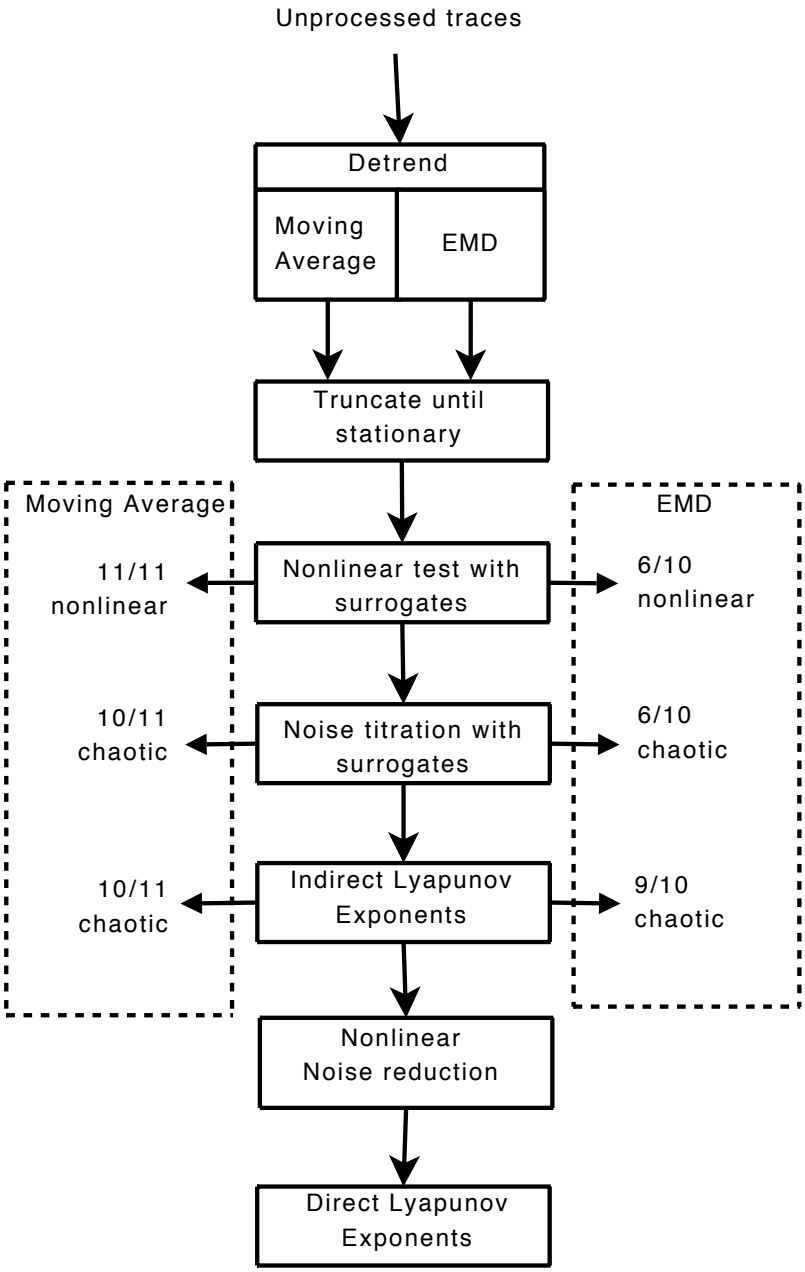


Figure 12: Flowchart of the tests run to gather evidence for chaos. A summary of results on the left of the figure are after processing with a moving average. The summary of results on the right are after detrending with Empirical Mode Decomposition (EMD).

we take this as evidence that the divergence is more likely due to deterministic chaos rather than stochasticity (Figure 10).

Nonlinearity, Noise Titration and Lyapunov Exponents

Evidence of chaos was suggested in the majority of traces (16 out of 21) using a noise titration with the surrogates nonlinear test (Table 2). Applied to linear autoregressive (AR) models fitted to the experimental data, this test correctly identified forty true negatives and only two false positives. In some cases, the results of the experimental time series vary depending on the method of detrending, with some (4 out of 10) EMD detrended time series failing the test for nonlinearity.

The nonlinear surrogates test exhibits a length dependence and so the shorter time series failed (Table 2). The nonlinear predictability was computed for a long time series that was steadily truncated to provide a comparison of p-value against series length (Figure 13). The p-values do not consistently indicate nonlinearity for times shorter than 400 samples.

An indirect method for maximal Lyapunov exponent calculation that fitted deterministic models to the Ca^{2+} time series, gave positive exponents for all experimental time series except for Nod3 and Nod4. All negative controls correctly gave negative maximal Lyapunov exponents.

Since the majority of the traces passed a test for nonlinearity the system can be considered nonlinear, justifying the application of nonlinear noise reduction techniques. Once the experimental Ca^{2+} spiking traces were noise-reduced, a direct Lyapunov calculation method was performed. The logarithm of the divergence of neighbouring points in phase space against time revealed a clear linear trend in the majority of the time series, indicating exponential divergence. This is shown in Figure 14. Taking an average gradient gave a Lyapunov exponent of $0.014s^{-1}$ for time series detrended using a moving average and $0.013s^{-1}$ for time series detrended using EMD.

2.5.2 *Tests for Comparison*

Nonlinear System with Random Interspike Intervals

Properties of the autocorrelation of interspike intervals have been used to support the idea of stochastic spike activation in four cell types from mice and humans [119]. In order to compare our initial results, which tend to support the case for determinism, with the stochastic

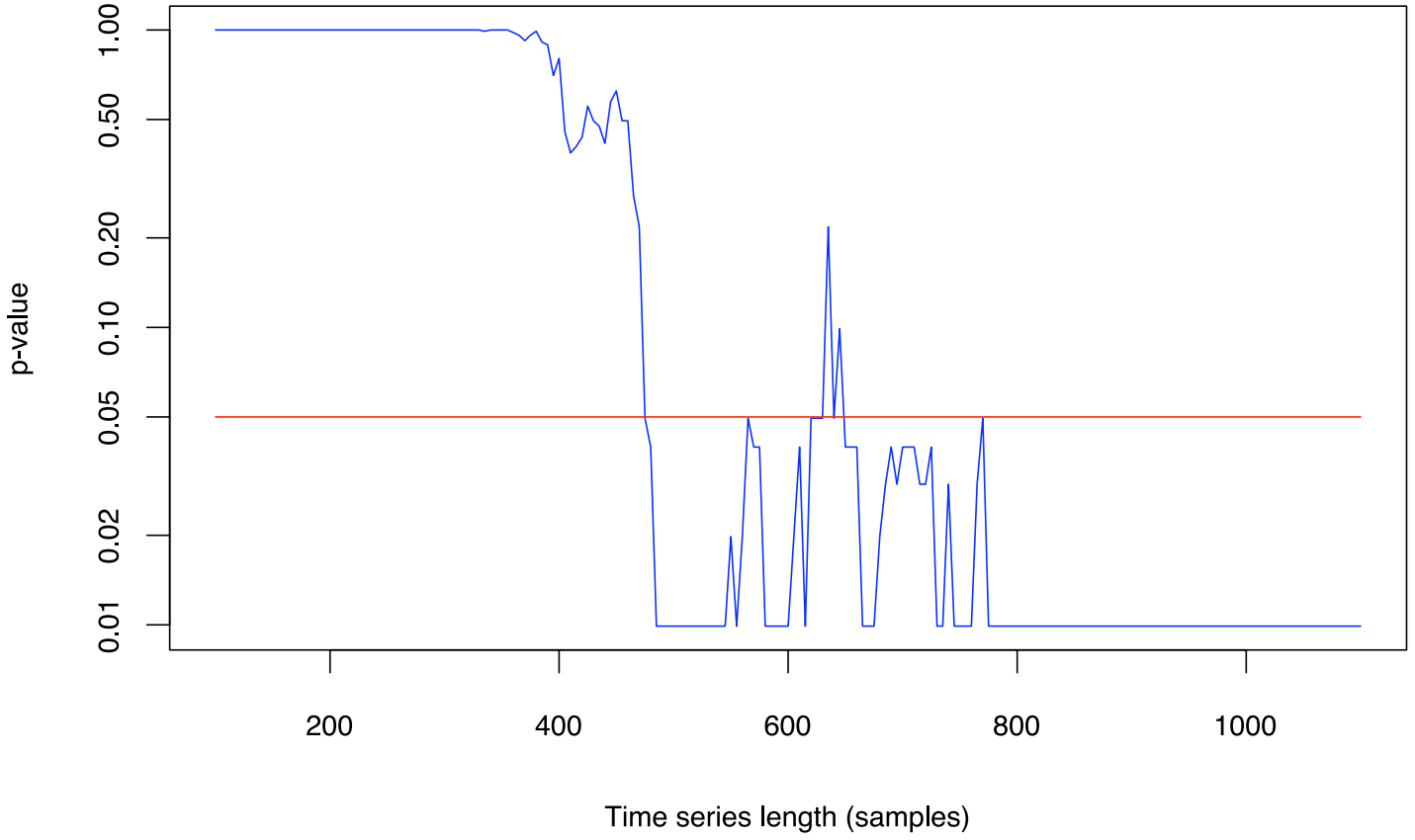


Figure 13: Semi-log plot of p-value against series length for a single time series that passes the zeroth surrogates test for nonlinearity. Each p-value was calculated using 100 surrogates. Signs of nonlinearity are not detected until the length of series being tested is greater than 400 samples (the sample time is 5 seconds). P-values do not drop to show significant nonlinearity, which is marked with a red line, until the time series is approximately 500 samples long.

Table 2: Stationary time series of Ca^{2+} concentration with lengths given by the number of samples. The sample time is 5 seconds. The spikes column indicates the number of spikes in the time series. P-values are given for the null hypothesis that each nod factor time series was generated by a linear process. The p-values are calculated by running the zeroth surrogates test, which was also used for a noise titration to get a limit for the noise that could be added without destroying evidence of nonlinearity.

Time Series	Detrending	Length	# Spikes	Zeroth	Noise Titration %
Nod1	MA	500	31	0.01	16
	EMD	500	34	0.01	16
Nod2	MA	400	36	0.01	25
Nod3	MA	700	45	0.01	23
Nod4	EMD	600	46	0.02	20
Nod5	MA	480	30	0.01	20
	EMD	741	46	0.01	12
Nod6	MA	339	14	0.01	0
	EMD	359	15	0.20	0
Nod7	MA	1058	46	0.01	9
	EMD	1170	50	0.01	11
Nod8	MA	1100	55	0.01	15
	EMD	1029	47	0.01	0
Nod9	MA	409	15	0.01	1
	EMD	260	9	0.83	0
Nod10	MA	400	22	0.01	20
	EMD	520	30	0.01	16
Nod11	MA	440	46	0.04	24

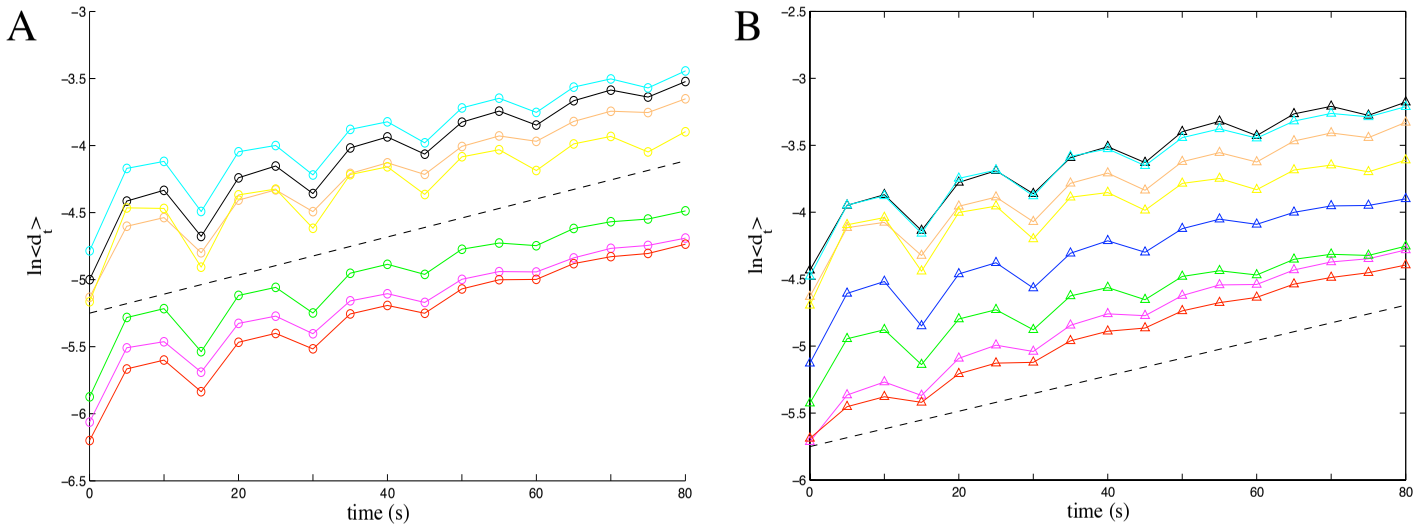


Figure 14: Semi-log plot of the average divergence <dt > of the nearest neighbors of each point in the time series as a function of time. The short term fluctuations are due to the periodicity of the signal, but the average distance clearly grows exponentially. The data points pictured are for an embedding dimension of seven, and consistent values for the maximal exponent are achieved once the dimension is greater than or equal to six. The exponent is 0.0142 for traces detrended with a moving average (A), and 0.0132 for EMD (B). These values are given by the slope of the black dashed lines. For each trace, the exponent was computed as the average of three slopes: 1) the slope through local maxima; 2) the slope through local minima; and 3) the slope through the average of local minima and maxima. The final value of the exponent was computed by averaging over all traces. Computations were not done for traces Nod6 and Nod9 for either detrending because of the short series length. We remark that the same computations for the stochastic spiking model give a semi-log plot of the form $< d_t > \sim t^{\frac{1}{\alpha}}$, indicating diffusive-like divergence.

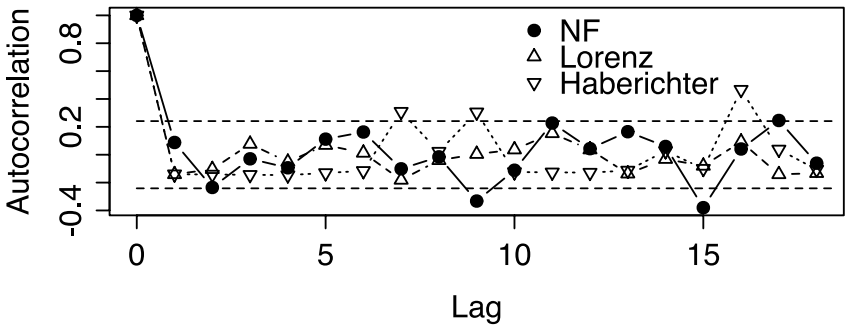


Figure 15: Autocorrelation of cycle times for a Nod Factor time series and two positive controls based on chaotic mathematical models. The X axis is the lag measured in number of samples (sample time is 5 seconds). All time series show a rapid drop to within the 95% confidence interval for white noise which is marked with horizontal dashed lines and represents no identifiably repeating patterns.

hypothesis, the autocorrelation of the intervals between maxima was calculated for two known chaotic differential equations and an experimental Nod factor Ca^{2+} spiking time series. For a purely random time series (white noise) the autocorrelation is close to zero. This is depicted in Figure 15 in which horizontal dashed lines mark the approximate 95% confidence interval for white noise. This confidence interval is calculated as $\pm\frac{1.96}{\sqrt{N}}$ where N is the length of the series of interspike intervals. Both the mathematical models and the experimental data show a rapid drop in autocorrelation indicating that successive intervals are not correlated. However, the mathematical models act as positive controls revealing that the drop in autocorrelation is not necessarily down to stochastic effects.

It must be pointed out that nonlinear time series analyses cannot provide a definite answer regarding the nature of spike activation and interspike times in the system. We considered a nonlinear deterministic model for the spike waveforms, with randomly-chosen interspike intervals. As expected, this signal clearly appears nonlinear; however, it also appears chaotic using a noise titration. This demonstrates that some conventional tests used to detect chaos are unable to discern between purely chaotic systems and a carefully designed deterministic spiking system with stochastic activation. For this reason we use a number of different techniques with the goal of presenting as thorough an analysis of the experimental Ca^{2+} oscillations as possible. A direct Lyapunov calculation for the time series with stochastic interspike times does not exhibit a clear exponential divergence. Figure 14 shows the divergence to be of the form $t^{\frac{1}{\alpha}}$, characteristic of a randomly

perturbed deterministic system. The indirect method also indicates that the majority of models fitted to the time series with stochastic interspike intervals have a negative Lyapunov exponent.

Determinism

The results of determinism tests are somewhat subjective and were therefore not used to support our conclusions. In contrast, the findings from one such test have been used as evidence of stochasticity in Ca^{2+} oscillations [97]. All traces obtained from our experiments pass the three statistical tests for determinism proposed by Aparicio [3] without the use of noise reduction. A negative control using random numbers fails the three determinism tests.

We evaluated the Kaplan and Glass measure for determinism on the Lorenz system and the Haberichter chaotic model of Ca^{2+} oscillations [39], both with 10% noise added to mimic the noise estimated in the experimental data. This method is based on a vector reconstruction of the attractor over a grid of boxes. A determinism factor, $\bar{\Lambda}$, is calculated where $\bar{\Lambda} = 1$ indicates full determinism and $\bar{\Lambda} = 0$ indicates complete randomness. The Lorenz time series had a determinism factor of $\bar{\Lambda} = 0.88$, and $\bar{\Lambda} = 0.78$ for the chaotic spiking model. Both of these time series would be classed stochastic using the criteria from other studies which required $\bar{\Lambda} > 0.9$. The Ca^{2+} spiking time series in this study have a low determinism factor, $\bar{\Lambda} < 0.3$. These results show the limitations in using only one metric to characterize noisy data sets.

2.6 DISCUSSION

Chaos is common in nature. For instance, the gravitational three body problem can exhibit deterministic chaos and numerous further examples exist for which chaotic behaviour has been identified or suggested, ranging from the solar system, weather, population dynamics, to Brownian motion and diffusion [32]. An interesting example of the potential relevance of chaotic flexibility has been discussed for human heart beats. It has been suggested that normal heart behaviour might be chaotic and can thus respond efficiently to perturbed conditions, whereas diseased hearts are more stable in their frequencies and less able to make necessary adjustments to stress [102]. However, chaos may also be involved in the destabilisation of heart rhythms, as quasiperiodicity and intermittency have been observed in the Ca^{2+} oscillations of cultured cardiomyocytes degenerating into chaos-like behaviour that would be fatal in-vivo [11]. Whether or not biological

systems such as the heart or brain are really chaotic is still the subject of much debate and on-going research.

Using a range of techniques from nonlinear time series analysis we have gained some evidence suggesting that the Ca^{2+} spiking in the root hair cells of M. truncatula might be chaotic. We first demonstrated that the majority of the time series show the Ca^{2+} oscillations to be nonlinear. To check for false positives we also tested linear models fit to the experimental data. The two false positives we obtained show that the test for nonlinearity can be fallible in some cases, should not be considered absolute, but nevertheless provides evidence of nonlinearity. We then performed a test for chaos using the noise titration technique. This test indicated that the majority of the Ca^{2+} time series were nonlinear in the presence of additive noise. This can be viewed as evidence of chaos [101]. Although the majority (19 out of 21) of the negative controls were correctly identified, the two false positives from the nonlinear test also passed the noise titration. Furthermore, a synthetically produced time series consisting of deterministic spikes separated by stochastic interspike intervals - a model that has been proposed for Ca^{2+} oscillations in animal systems - was also classed as chaotic by the noise titration method. Although this model is largely deterministic and nonlinear, it is not chaotic. This demonstrates that classification using the noise titration method should be done with caution.

Using a combination of an indirect method to compute the probable sign of Lyapunov exponents and a direct method to calculate the magnitude and type of the divergence, evidence of chaos was revealed in the Ca^{2+} oscillations and controls without any false positives. To our knowledge, this particular combination of direct and indirect methods of Lyapunov exponent calculation with the use of controls has not been used before.

In animals, the hypothesis that Ca^{2+} oscillations, experimentally obtained from hepatocytes, originated from a deterministic system was rejected [97]. The conclusion that these oscillations are "prevalently stochastic" was reached because one of two time series failed a nonlinear test, and the one that passed had a determinism score of $\bar{\Lambda} < 0.9$ as provided by a Kaplan-Glass analysis. We have given two examples of chaotic oscillations that fail to meet this criterion under similar noise conditions to the experimental data being considered. The noise present in our experimental data (around 10%), results in some of the individual tests producing inconclusive answers, but the combination of all results presents a stronger case which suggests

that the oscillations in M. truncatula are produced by a nonlinear, deterministic system.

In order to understand the fundamental nature of seemingly erratic calcium oscillations, the question of randomness or chaos arises and needs to be sufficiently addressed. To indisputably demonstrate stochasticity as a main driving mechanism in calcium oscillations, determinism must be eliminated. This is a non-trivial task for a number of reasons. Fundamentally, given that noise is nearly always present and the high demand on data quality and quantity for most non-linear techniques to work robustly, this distinction between stochasticity, noise, and low-dimensional chaos can rarely be achieved. Practically, the choice of parameterisation is often known to be approximate and deviations are called stochastic effects within the chosen framework and reduced phase space. However, there now exists a wealth of advanced tools and approaches from time series analyses and dynamical systems theory, which can be employed to shed light on the nature of experimental data and offer possible interpretations. In accepting randomness too readily, the exciting discovery of a biological system taking advantage of attributes of chaotic motion would be missed and some of its most interesting features labeled as chance occurrences.

A number of properties of dissipative chaotic systems make them suitable for Ca^{2+} signaling. First, and perhaps counterintuitively, a theoretical study on Ca^{2+} oscillations has shown that both the sensitivity to parameter perturbations and the capacity to attune to a forcing frequency do not depend on the oscillations being chaotic or regular [96]. This means that, despite the sensitive dependence on initial conditions, chaotic systems can be equally robust and flexible as regular systems in a highly variable biological environment. While these statements are based upon evidence from a particular model, they can be generalized.

In non-conservative systems, chaotic trajectories are restricted to lie upon either strange attractors or chaotic saddles. These two cases represent sustained or transient chaos, respectively. The saddles are hyperbolic, and as such they are structurally stable and deform smoothly with parameters. Moreover, it has been shown that the transient time spent tracing a chaotic saddle changes slowly with increasing levels of noise [69]. The case of sustained chaos is similar: strange attractors typically retain their shape regardless of small parameter perturbation (except at crisis values). Thus the trajectories that trace the attractors also maintain their characteristic shape in noisy environments. The

consequence is that the patterns made by system observables — here the oscillating Ca^{2+} level — can be robust despite fluctuations.

These qualities are advantageous to the symbiosis signaling pathway under study, which has been shown to take part in two important symbioses that are evolutionary separated by hundreds of millions of years [65]. Each symbiosis leads to a different Ca^{2+} oscillation signature despite the use of common components. The existence of the multiple steady states excludes the possibility that the signaling pathway is a stationary, linear process.

The possibility that the system jumps from one attractor to another in response to different input signals would have important implications, but the capacity for dual signal generation could also be a sign that the system is controlling chaos, i.e. the two signals represent subregions of a larger chaotic set. The control of chaotic motion, as originally proposed [93], utilizes that the state of the system visits any neighborhood of periodic orbits of every period. Tiny controlling effects can then be adeptly used to direct the behaviour to any periodic motion. The concept has been widely used in circuitry, lasers, chemistry, low-energy orbit design, and even to direct the rhythms of the heart. In the case of Ca^{2+} oscillations one candidate for the source of the pertubations is Ca^{2+} influx [120]. The control algorithm is attractive because of its efficiency, and could be used here to maintain the periodicity of the oscillations, to synchronize spatially separate components, or to specify one of the two signals.

It remains to be discovered whether chaos control is being harnessed for the efficient tuning of the Ca^{2+} oscillations or if chaotic flexibility is an essential factor for signaling specificity. Discovering further examples of nonlinearities and chaos within the cell would have implications for the way we view the principles of signaling pathways. One reason to suspect intracellular chaos is simply that it can be produced rather easily by relatively few strongly-interacting components, and it is common in many natural systems. As has been shown for noise, biological systems are capable of using common effects to its advantage. Given that chaotic systems can indeed be robust, and that chaos control enhances adaptability to environmental changes at less energetic costs and with accurate targeting of desired behaviour, we find this a fascinating speculation for biological signaling. It may come as no surprise to learn that evolution could have beaten physicists to the discovery that small perturbations can be efficiently used to control chaotic systems [41, 116].

3

IDENTIFYING MODELS FROM EXPERIMENTAL DATA

3.1 OVERVIEW

This chapter looks at ways of analysing a chaotic time series to identify the system that produced the time series. The investigation was done using a set of known chaotic equations and attempts were made to recover the equations from data they produced. It was hoped that this investigation would lead to a method to identify unknown components in the Ca^{2+} spiking system. However, despite suggesting and benchmarking some useful techniques, this investigation was not successful in producing a framework that could be applied to experimental data.

3.2 INTRODUCTION

As discussed in Section 1.2.4, the components in *Medicago truncatula* root hairs that contribute to Ca^{2+} oscillations are not well understood. Gaps in our knowledge about the system will translate to many modelling iterations where different hypotheses are modelled and then checked against the available experimental data. These Ca^{2+} time series contain information about the system that will guide the modelling process. This chapter looks at methods to computationally extract information from the time series to produce mathematical models with minimal manual intervention.

3.2.1 *System Identification*

In this work, system identification is defined to be the automated or semi-automated production of a mathematical model from a set of experimental data. This problem has been investigated in separate fields of research using various methods and sometimes using different nomenclature.

One approach to system identification is the generation of linear differential equations from a time series of variables [74]. The techniques used for the generation of linear differential equations will

not produce nonlinear differential equations without including prior information about the physical system [74].

If no prior information is available, a Volterra-Wiener series can be fitted to the data using an Orthogonal Algorithm proposed by Korenberg [64, 74, 129]. The Volterra-Wiener series, $x_0 \rightarrow x_N$, is assumed to have a memory κ, degree d, number of terms M and parameters $a_{0..M-1}$:

$$\begin{aligned} x_n = {} & a_0 + a_1 x_{n-1} + \cdots + a_\kappa x_{n-\kappa} \\ & + a_{\kappa+1} x_{n-1}^2 + a_{\kappa+2} x_{n-1} x_{n-2} + \cdots + a_{\kappa+\kappa} x_{n-1} x_{n-\kappa} \\ & + \cdots + a_{2\kappa+1} x_{n-2}^2 + \cdots + a_{M-1} x_{n-\kappa}^d \,. \end{aligned} \tag{3.1}$$

With a degree of 1, the Volterra-Wiener series corresponds to an AR model (Section 2.4.1). A Volterra-Wiener series has the potential to make good predictions but it is a "black box" model that gives little insight into the physical components that make up a system.

In the field of genetic programming, the automatic production of equations from data is given the name "symbolic regression". Genetic programming has successfully been used to recover differential equations for simple and chaotic pendulum systems [111].

System biologists sometimes describe the problem of system identification as "network building" and work in this area has produced equations in the form of nonlinear S-systems [61, 90]. An S-system is a collection of equations with the following structure:

$$\frac{dX_i}{dt} = \alpha_i \prod_{j=1}^{n} X_j^{g_{ij}} - \beta_i \prod_{j=1}^{n} X_j^{h_{ij}} \,, \tag{3.2}$$

where n is the number of state variables ($X_{i\ldots n}$) and α_i, β_i, g_{ij} and h_{ij} are parameters. An S-system is flexible enough to describe a gene network, but because the structure of the system is set prior to system identification, fitting a model to experimental data is a tractable parameter estimation problem.

There have also been successful attempts in system biology to infer nonlinear ordinary differential equations (ODEs) [109, 47, 87, 88] from experimental data.

For a system identification method to be used on the symbiotic Ca^{2+} oscillations in *Medicago truncatula*, the method must support the generation of multidimensional equations from time series describing only a single variable. The complete system state cannot be used since only time series of Ca^{2+} concentration are available. In fact, the variables that make up the system state are the subject of speculation. The Ca^{2+} time series passes tests for chaos and so the system identification

method must also be able to produce nonlinear and chaotic system equations. The aim of this research is to produce a mathematical model of Ca^{2+} oscillations that will enhance our understanding of the biological system being modelled. A system of nonlinear ODEs is the best candidate for a model with a biological interpretation since the majority of existing models for Ca^{2+} oscillations consist of nonlinear ODEs. None of the studies previously mentioned meet all 3 of our requirements. This encouraged us to investigate a different approach. We suggest a general strategy for system identification inspired by Sakamoto and Iba [109], break this strategy into subproblems and investigate the subproblems with regard to chaotic systems.

3.2.2 *Exploring Model Space*

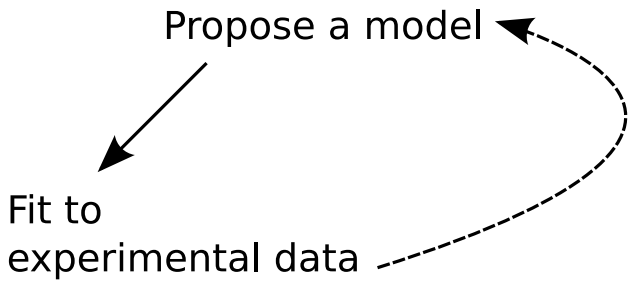


Figure 16: A rough sketch of how an automated system can be used to generate a mathematical model for a poorly understood system.

A set of mathematical models that could be used to describe the system can be thought of as a model space. The act of trying different models to see if they are plausible can be thought of as exploring the model space. If exploring model space is done by evaluating models manually, the number of models attempted will be relatively small i.e. only a small part of model space will be explored due to time constraints. Although the models attempted by a human modeller will likely be both parsimonious and plausible, the reduction in the volume of model space being searched makes it less likely to find a good model and reduces the probability of finding the best model that meets specified criteria.

This chapter gives a piece by piece description of some methods that could possibly be used to explore model space in an automated way. A possible shortcoming of these approaches is that the models generated in an automated approach may not be as sensible or as plausible as those proposed by human experts. A great advantage is

the throughput of automated methods which can examine hundreds of models in the time required to analyse a single model by hand.

A rough outline of the methods is given in Figure 16. An algorithm proposes a model which is then tested against experimental data. Information on how well this model, and previous models, describe the experimental data is then considered when proposing a new model. Using such a strategy should result in model proposals that improve with each iteration.

3.3 PARAMETER ESTIMATION

A mathematical model based on Ordinary Differential Equations (ODEs) contains variables that change over time and parameters which are considered to be fixed values. In an intracellular model, typical variables are the concentrations of proteins and ions, while parameters can represent the values of rate constants and volumes.

Ideally all parameters should be measured from the system being modelled. However, for the majority of intracellular systems, background knowledge is typically incomplete and parameters will have to be estimated. Estimating parameters is a critical operation since a model will have qualitatively different behaviour for alternative parameter values. One way to estimate parameters is the ask the question, "for this model, what are the parameter values (within realistic ranges) that will give me a behaviour most similar to the available experimental data ?"

Asking this question of the parameters also yields a useful measure of how well a given model could possibly account for experimental results. Parameter estimation has been used in such a way to propose the gene GIGANTEA as a component of the circadian clock in *Arabidopsis thaliana* after parameter estimates with an existing model failed to account for experimental data [75].

Although well established with many real world applications, parameter estimation is a research topic in its own right. Parameter estimation on chaotic systems is more complex, as illustrated in section 3.3.1, and requires novel algorithms with additional complications. To our knowledge, parameter estimation on a chaotic model using noisy intracellular experimental data has never been attempted. Because of this some numerical experiments were performed to assess the suitability of chaotic parameter estimation as a way of assessing models in an automated modelling system.

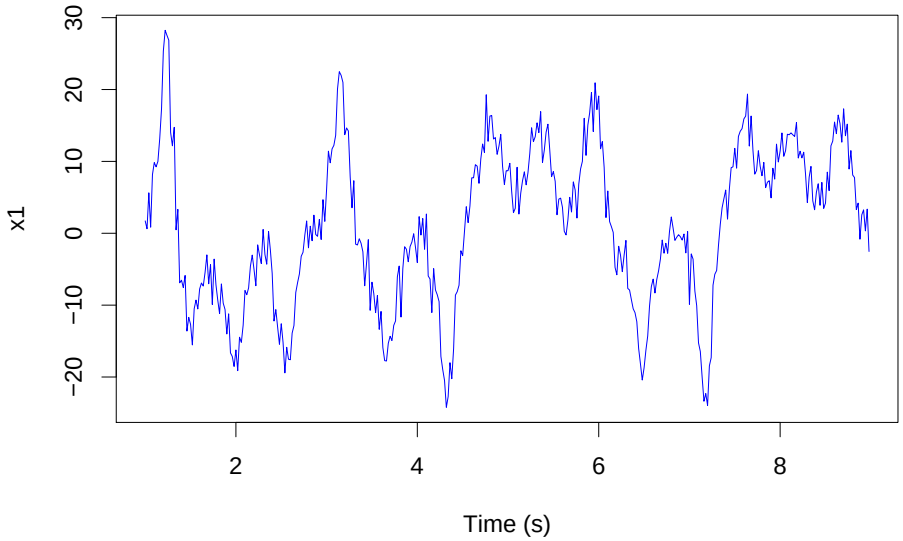


Figure 17: Simulated experimental data produced by adding 20% noise to a chaotic solution of the Lorenz equations.

Simulated experimental data was produced from the Lorenz system of equations [76]:

$$\dot{x}_1 = -\lambda_1 x_1 + \lambda_1 x_2 \tag{3.3}$$

$$\dot{x}_2 = -\lambda_2 x_1 - x_2 + x_1 x_3 \tag{3.4}$$

$$\dot{x}_3 = -\lambda_3 x_3 + x_1 x_2 \, . \tag{3.5}$$

Using the parameter values $\lambda_1 = 10$, $\lambda_2 = 46$ and $\lambda_3 = 2.67$, the equations were integrated to produce a time series. Twenty percent Gaussian noise was then added to create a noisy chaotic time series (Figure 17).

Parameter estimation algorithms were assessed by how accurately they were able to recover the parameter values used to generate the simulated experimental data.

3.3.1 *Single Shooting*

The single shooting algorithm is the most commonly used algorithm to perform parameter estimation on non-chaotic systems. It is described here as an introduction to parameter estimation and to illustrate the algorithm's shortcomings when working with chaotic data.

The single shooting algorithm is a combination of a 'cost' or 'fitness' function with an optimiser.

Cost Function

The cost function produces a numeric value indicating how well the model with a given vector of initial conditions and vector of parameters fits some experimental data. In this text, the cost function is assumed to give lower numeric values to indicate a good fit to the data and higher numeric values to indicate a poor fit to the data.

For ease of implementation and low computing cost the single shooting methods described here rely on the least squares cost function. A least squares estimate, $C_{LS}(\mathbf{x}_0, \boldsymbol{\theta})$, for a vector of initial conditions, $\mathbf{x}_0$, and a vector of parameters, $\boldsymbol{\theta}$, is the squared difference between a model, $\mathbf{x}_{i+1} = f(\mathbf{x}_i, \boldsymbol{\theta})$, observed through a measurement function, h, and a time series of data, $\{z\}_{i=0}^{M}$:

$$C_{LS}(\mathbf{x}_0, \boldsymbol{\theta}) \quad = \quad \sum_{i=0}^{M-1} [z_{i+1} - h(f(\mathbf{x}_i, \boldsymbol{\theta}))]^2 \, . \tag{3.6}$$

Despite its wide use in the estimation of parameters for both linear and nonlinear systems, least squares is not optimal when estimating the parameters of a noisy nonlinear system [80]. However, as shown below and in [84] the results are a good approximation when dealing with nonlinear systems.

There has also been practical success for estimating the parameters of oscillating systems in plants using a custom cost function based on qualities of the experimental data such as period, phase, broadness of peak and amplitude [75]. An advantage of this method is that it is possible to make the cost function simpler to optimise and that the cost of estimated parameter values close to the actual values are low and vary smoothly. One disadvantage of a custom cost function is that the cost function is dependent on the experimental data — for instance successful results with the Lorenz equations will not be easily translated to other systems since they depend on the custom cost function. Another issue is that such a cost function requires trial and error to design correctly which increases the development time of a parameter estimation method. Also, as shown in section 3.3.1 modern global optimisers are able to make good parameter estimates despite issues with least squares.

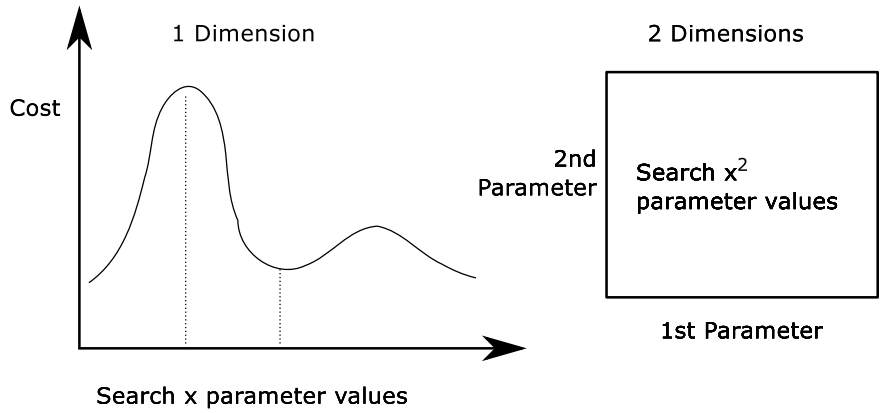


Figure 18: As the number of dimensions, A, of parameter space increases the number of points required for an exhaustive search is x^A where x is the number of parameter values checked in each dimension.

Optimisers

A system with a number of parameters, A, can be thought of as having an A dimensional parameter space. Each point in the parameter space will have a cost, C, associated with it which can be calculated using a cost function. Estimating parameter values can be thought of as exploring parameter space to find a set of parameters θ_{min} that have the lowest cost. A section of parameter space for the simulated experimental data generated from the Lorenz system is shown in Figure 19.

This problem as described above is a global optimisation problem and is best performed by an optimisation algorithm. It is not realistic to exhaustively search parameter space as the number of parameters increases > 3 (Figure 18).

Global optimisation is non-trivial because it is not possible to know if a set of parameters on a local minimum, θ_{local}, are globally minimum. Information about the gradients of the parameter space are typically used by a class of optimisers that are deterministic. These deterministic optimisers give reproducible results and will always take the same route through parameter space to an estimated set of parameters that will be on a local minimum. Another class of optimisers, known as stochastic optimisers, don't usually use gradients and have random elements in their operation which leads to unrepeatable and varying solutions. Stochastic optimisers have been inspired by examples in nature such as birds swarming [43] or evolution [10] or have designed around heuristic behaviour [106]. Generally, stochastic optimisers don't even guarantee that the parameters chosen even lie on a local minimum of the cost function.

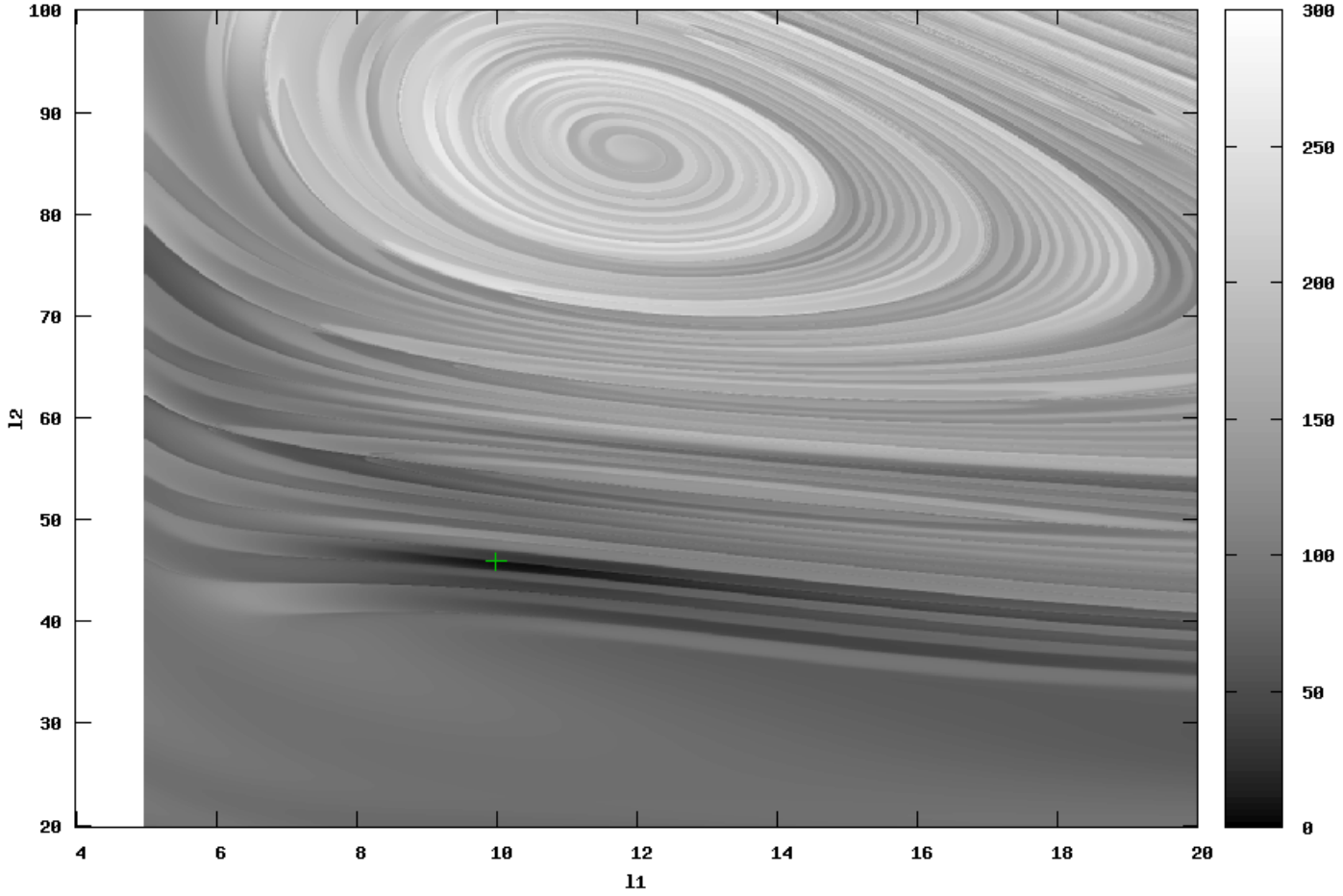


Figure 19: A 2D slice through the 3D cost landscape of least squares fitting of the Lorenz equations with 2 of the 3 parameters unknown (λ_1 labelled l1, and λ_2 labelled l2). The initial conditions are known. The green cross marks the actual parameter values (l1 = 10 and l2 = 46). Dark shading indicates a good fit of the model to the simulated experimental data with lighter shades denoting poorer fits.

Although not highly regarded in some circles, it has been shown that stochastic global optimisers outperform deterministic optimisers on nonlinear parameter estimation problems. The Stochastic Ranking Evolutionary Strategy (known as SRES and described in [108]) has been shown to make particularly good parameter estimates [84].

The SRES algorithm was used during this study and found to perform well in the majority of cases. However, it was found that, for some optimisation problems, the SRES algorithm was not as robust or as scalable as a particle swarm algorithm. Because of this, a particle swarm optimiser is used in some parts of this investigation. The particle swarm implementation was developed from a description by He et al. [43]. A modification was made where the initial parameter values were obtained from deterministic Sobol sampling [103] which generates a more even sampling of parameter space when compared to using a uniform random distribution. The modified particle swarm optimiser is shown in Algorithm 1. It was found that the Sobol modification improved the performance of the algorithm for larger dimensional problems.

The Single Shooting Algorithm

A form of the simple shooting algorithm is given in Algorithm 2 for an arbitrary cost function, denoted C, and an arbitrary optimiser.

Since single shooting is searching for both initial conditions, $\mathbf{x}_b$, and parameter values, θ_b, the search space is higher dimensional, and possibly more complex, than parameter space alone. An example of this is shown in Figures 20 and 21 which also illustrate a problem with using single shooting on chaotic systems. Because chaotic systems are sensitive to initial conditions (as described in section 2.4.8), a small change in the initial conditions will result in a large change to the cost function. This sensitivity to initial conditions is not just a practical problem and makes precise estimation of initial conditions theoretically impossible with the limited precision of floating point numbers in digital computers [54].

It is impossible to accurately estimate initial conditions and hence the true values of the parameters of chaotic system using single shooting. However, it is sometimes feasible to find a trajectory that shadows the experimental data for hundreds of datapoints (Figure 22).

Algorithm 1 Particle Swarm Optimiser

Require: N_e, the number of elements in a particle
Require: N_p, the number of particles to use
Require: N_g, the number of generations to run
  $\mathbf{v}_i$, the velocity of particle i
  Create particles, $\{\mathbf{P}\}_{i=1}^{N_p}$, by Sobol sampling
  $\mathbf{r} \in \mathbb{R}^{3\times N_e}$, random numbers $0 < r_{ij} < 1$
  for $i = 1 \rightarrow N_p$ **do**
    for $j = 1 \rightarrow N_e$ **do**
      $v_{ij} \leftarrow 0.2 r_{1j} P_{ij}$
    end for
  end for
  O_i, the objective function value of particle i
  O_i^b, the best objective function value of particle i
  $\mathbf{P}_i^b$, the particle position where O_i^b was obtained
  O^b, the global best objective function value
  $\mathbf{P}^b$, the particle position where O^b was obtained
  inertia $\leftarrow 0.9$
  for $g = 1 \rightarrow N_g$ **do**
    for $i = 1 \rightarrow N_p$ **do**
      Calculate O_i
      if $O_i < O^b$ **then**
        $\mathbf{P}^b \leftarrow \mathbf{P}_i$
        $O^b \leftarrow O_i$
      end if
      if $O_i < O_i^b$ **then**
        $\mathbf{P}_i^b \leftarrow \mathbf{P}_i$
        $O_i^b \leftarrow O_i$
      end if
      for $j = 1 \rightarrow N_e$ **do**
        $v_{ij} \leftarrow \text{inertia} \times v_{ij} + 2r_{2j}(P_{ij}^b - P_{ij}) + 2r_{3j}(P_j^b - P_{ij})$
      end for
      $\mathbf{P}_i \leftarrow \mathbf{P}_i + \mathbf{v}_i$
    end for
    inertia $\leftarrow$ inertia $- (0.9 - 0.4)/N_g$
  end for
  Return $\mathbf{P}^b$

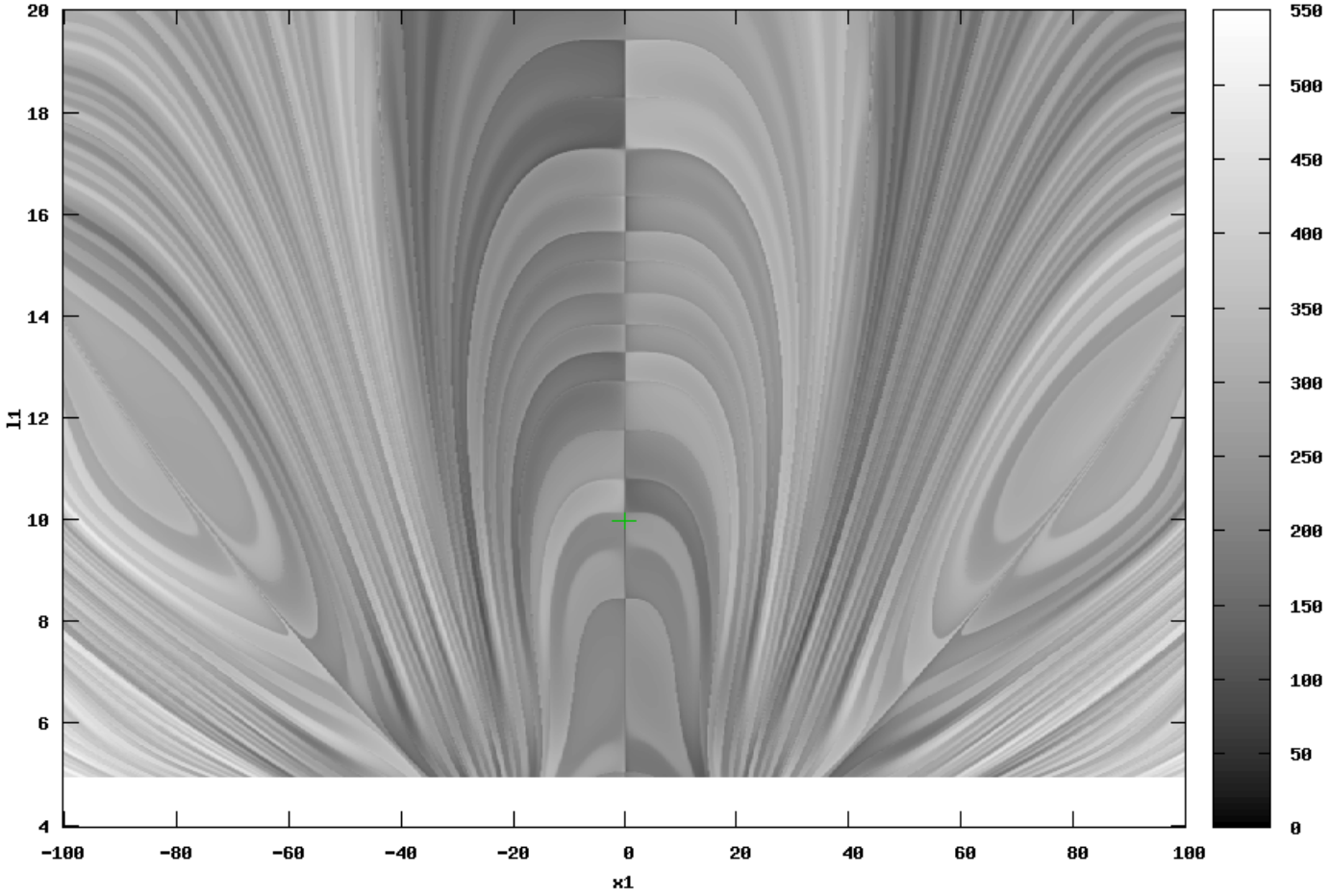


Figure 20: A 2D slice through the 3D fitness landscape of least squares fitting of the Lorenz equations with 1 of the 3 parameters unknown (λ_1 labelled l1) and 1 of the initial conditions unknown (x_1 labelled x1). The green cross marks the actual parameter and initial condition values (l1 = 10 and x_1 = 0.0001). Dark shading indicates a good fit of the model to the simulated experimental data with lighter shades denoting poorer fits.

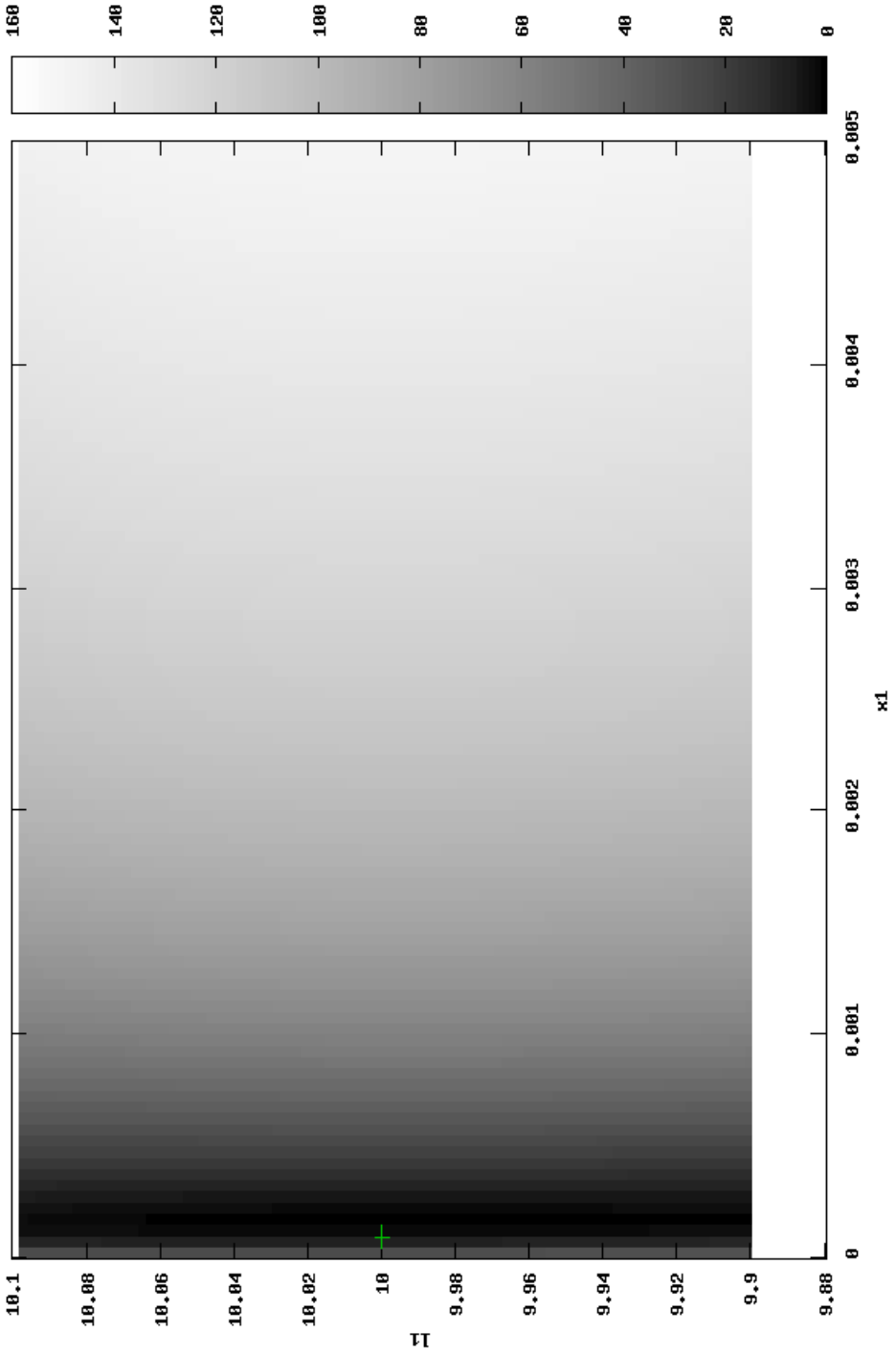


Figure 21: A zoom in on the area around the green cross in Figure 20

Algorithm 2 Single Shooting

1: **for** $i = 1 \rightarrow M$ **do**
2: Initialise best parameters θ_b and best initial conditions $\mathbf{x}_b$
3: Optimiser selects parameters θ_i and initial conditions $\mathbf{x}_{0i}$
4: **if** $C(\mathbf{x}_{0i}, \theta_i) < C(\mathbf{x}_b, \theta_b)$ **then**
5: $\mathbf{x}_b \leftarrow \mathbf{x}_{0i}$
6: $\theta_b \leftarrow \theta_i$
7: **end if**
8: Update optimiser with $C(\mathbf{x}_{0i}, \theta_i)$
9: **end for**
10: Return $\mathbf{x}_b$ and θ_b

3.3.2 *Bayesian Filters*

An alternative to shooting methods for estimating the parameters of chaotic systems was suggested by Sitz et al. [118] who described an application of the Unscented Kalman Filter (UKF). The UKF can be thought of as a type of Bayesian filter that makes a series of predictions over a time series. The filter updates its knowledge from actual time series measurements which should result in more accurate predictions as the filter moves along the time series (Figure 23).

Most Bayesian filters assume a state space model for the time series,

$$\mathbf{x}_{t+1} = f(\mathbf{x}_t, \theta) + \boldsymbol{\varphi} \tag{3.7}$$

$$y_{t+1} = h(\mathbf{x}_{t+1}) + \eta \tag{3.8}$$

where $\mathbf{x}_t$ is a vector describing the state of the system at time point t, f is a function corresponding to the model of the system, θ are the parameters of the model, $\boldsymbol{\varphi}$ is a system noise term, y_t is a prediction of a scalar time series measurement, h is a measurement function and η is additive noise. Actual time series measurements are denoted here as z_t where $\{z\}_{t=0}^{M}$.

Algorithm 3 Using a Bayesian filter to make time series predictions

Require: estimated $\mathbf{x}_0$
1: **for** $t = 0 \rightarrow M - 1$ **do**
2: $\mathbf{x}_{t+1} \leftarrow f(\mathbf{x}_t, \theta) + \boldsymbol{\varphi}$
3: $y_{t+1} \leftarrow h(\mathbf{x}_{t+1}) + \eta$
4: update $\mathbf{x}_{t+1}$ using z_{t+1}
5: **end for**

The critical step in Algorithm 3 that distinguishes between different types of Bayesian filters is the update on line 4 and its use of the time

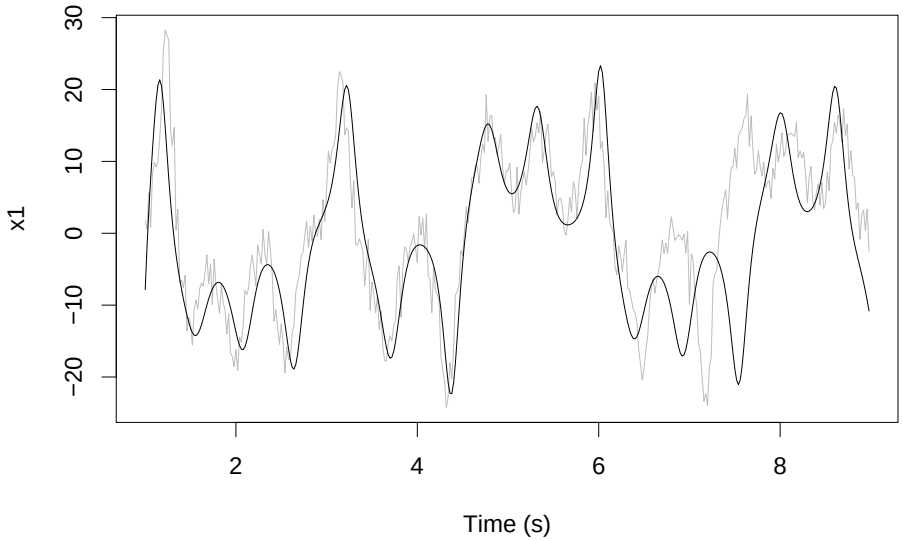


Figure 22: A fit of the Lorenz equations to the simulated experimental data using the SRES optimiser with unknown initial conditions. The simulated experimental data is given in gray and the fitted model in black.

series measurements z_{t+1} to improve knowledge about the internal state $\mathbf{x}_{t+1}$.

Algorithm 4 Using a Bayesian filter to estimate parameters

Require: estimated $\mathbf{x}_0$, θ_0
  for $t = 1 \rightarrow M - 1$ **do**
    $\mathbf{x}_{t+1} \leftarrow f(\mathbf{x}_t, \theta_t) + \boldsymbol{\varphi}$
    $y_{t+1} \leftarrow h(\mathbf{x}_{t+1}) + \eta$
    update $\mathbf{x}_{t+1}$ and θ_{t+1} using z_{t+1}
  end for

Bayesian filters can be modified to estimate parameters by making a small change as shown in Algorithm 4. This method changes the update step so that z_{t+1} is used to improve knowledge about both the state, $\mathbf{x}_{t+1}$, and the parameters θ_{t+1}. This is trivial to do in practice by making θ_{t+1} part of the state vector $\mathbf{x}_{t+1}$ i.e. treating the parameters as variables.

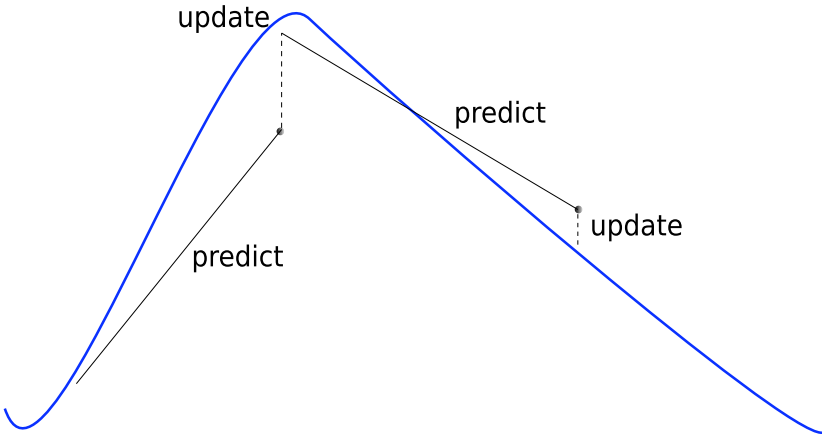


Figure 23: Cartoon of the operation of a Bayesian filter with filter steps given in black and a time series in blue. The filter makes a prediction and then updates its knowledge about the state of the system from a measurement. This updated state is then used to make a potentially improved prediction for the next data point.

Unscented Kalman Filter

The Unscented Kalman filter (UKF) has been proposed as a parameter estimation algorithm for nonlinear and chaotic systems [128, 118, 124]. This filter approximates the distribution of $\mathbf{x}_t$ as a Gaussian. This Gaussian distribution is propagated from $\mathbf{x}_t$ to $\mathbf{x}_{t+1}$ by only considering the mean and covariance. The mean and covariance of $\mathbf{x}_{t+1}$ and y_{t+1} are estimated using an Unscented transform. This transform deterministically samples an n-dimensional random variable using $2n+1$ weighted points known as sigma points [55].

An overview of the UKF has been given in a presentation by Merwe and Wan [81] labelled as Algorithm 2.1. It is reproduced here with additional comments. Some equations have been modified slightly for the special case, matching the situation of the symbiotic Ca^{2+} spiking system, where the time series measurements are scalar rather than vectorial.

An expected value for the initial state is set from prior information,

$$\hat{\mathbf{x}}_0 = \mathbf{x}_0 \,. \tag{3.9}$$

An initial state covariance matrix, $\mathbf{P}_0$, for an n dimensional variable is initialised as,

$$\mathbf{P}_0 = \varphi^2 \mathbf{I}_n \,, \tag{3.10}$$

where $\mathbf{I}_n$ is an $n \times n$ identity matrix and $\boldsymbol{\varphi}$ is a vector containing system noise standard deviations for each variable in the system (from Equation 3.7).

When analysing a time series, $\{z\}_{t=0}^{M}$, the following steps are applied to the time points $t = 1 \dots M$. First the matrix of sigma points, χ, that is used to sample the covariance, is calculated:

$$\chi_{t-1} = \left[\hat{\mathbf{x}}_{t-1},\ \hat{\mathbf{x}}_{t-1} + \gamma\sqrt{\mathbf{P}_{t-1}},\ \hat{\mathbf{x}}_{t-1} - \gamma\sqrt{\mathbf{P}_{t-1}}\right], \quad (3.11)$$

where each column in the matrix χ specifies a system state, $\mathbf{x}_{i,t-1}$, that will used as a sigma point to sample the next system state, $\mathbf{x}_{i,t}$. γ is a scalar defined as:

$$\gamma = \sqrt{n + \lambda} \quad (3.12)$$

$$\lambda = \alpha^2(n + \kappa) - n\,. \quad (3.13)$$

α is a controlling constant that determines the spread of sigma points around the mean and κ is a scaling parameter.

The sigma points are transformed through the nonlinear function representing the model of the system (from Equation 3.7),

$$\chi^*_{i,t|t-1} = f(\chi_{i,t-1}, \theta)\,, \quad (3.14)$$

where $i = 0 \dots 2n$, is the index of the sigma point being calculated. A weighted average of the transformed points, $\bar{\mathbf{x}}_t$, is then calculated:

$$\bar{\mathbf{x}}_t = \sum_{i=0}^{2n} W_i^{(m)} \chi^*_{i,t|t-1}\,. \quad (3.15)$$

$W_i^{(m)}$ is the weighting to use for point i,

$$W_0^{(m)} = \frac{\lambda}{n+\lambda} \quad (3.16)$$

$$W_i^{(m)} = \frac{1}{2(n+\lambda)} \quad \text{where } i = 1 \dots 2n\,. \quad (3.17)$$

Having calculated the weighted average, the covariance, $\bar{\mathbf{P}}_t$, can be estimated:

$$\bar{\mathbf{P}}_t = \sum_{i=0}^{2n} W_i^{(c)} \left[\chi^*_{i,t|t-1} - \bar{\mathbf{x}}_t\right] \left[\chi^*_{i,t|t-1} - \bar{\mathbf{x}}_t\right]^T + \mathbf{R}^{\mathbf{v}}\,. \quad (3.18)$$

The alternative weight, $W_i^{(c)}$, is given as:

$$W_0^{(c)} = \frac{\lambda}{n+\lambda} + (1 - \alpha^2 + \beta) \quad (3.19)$$

$$W_i^{(c)} = W_i^{(m)} \quad \text{where } i = 1 \dots 2n\,. \quad (3.20)$$

$\mathbf{R}^{\mathbf{v}}$ is the system noise covariance matrix:

$$\mathbf{R}^{\mathbf{v}} = \boldsymbol{\varphi}\boldsymbol{\varphi}^T\,. \quad (3.21)$$

The sigma points are then recalculated to take into account system noise,

$$\boldsymbol{\chi}_{t|t-1} = \left[\bar{\mathbf{x}}_t, \quad \bar{\mathbf{x}}_t + \gamma\sqrt{\bar{\mathbf{P}}_t}, \quad \bar{\mathbf{x}}_t - \gamma\sqrt{\bar{\mathbf{P}}_t}\right] , \tag{3.22}$$

and are then transformed through the measurement function (Equation 3.8):

$$Y_{i,t|t-1} = h(\boldsymbol{\chi}_{i,t|t-1}) . \tag{3.23}$$

A weighted average, $\bar{y}_t$, is then calculated to give an estimate for the time series measurement at time t:

$$\bar{y}_t = \sum_{i=0}^{2n} W_i^{(m)} Y_{i,t|t-1} . \tag{3.24}$$

This estimate is used to update the innovation variance [55]:

$$P_{\bar{y}_t,\bar{y}_t} = \eta^2 + \sum_{i=0}^{2n} W_i^{(c)} \left[Y_{i,t|t-1} - \bar{y}_t\right] \left[Y_{i,t|t-1} - \bar{y}_t\right]^T , \tag{3.25}$$

where η is the measurement noise introduced in Equation 3.8 which could be replaced with a covariance matrix if vectorial measurements were being considered.

A cross correlation matrix is also calculated at this stage:

$$\mathbf{P}_{x_t,y_t} = \sum_{i=0}^{2n} W_i^{(c)} \left[\boldsymbol{\chi}_{i,t|t-1} - \bar{\mathbf{x}}_t\right] \left[Y_{i,t|t-1} - \bar{y}_t\right]^T . \tag{3.26}$$

Information from the actual time series measurement, z_t, can then be incorporated into the filter:

$$\mathbf{K}_t = \mathbf{P}_{x_t,y_t} P_{\bar{y}_t,\bar{y}_t}^{-1} \tag{3.27}$$

$$\hat{\mathbf{x}}_t = \bar{\mathbf{x}}_t + \mathbf{K}_t(z_t - \bar{y}_t) . \tag{3.28}$$

To prepare for the next cycle an update is made to the state covariance:

$$\mathbf{P}_t = \bar{\mathbf{P}}_t - \mathbf{K}_t P_{\bar{y}_t,\bar{y}_t} \mathbf{K}_t^T . \tag{3.29}$$

At this point $\hat{\mathbf{x}}_t$ and $\mathbf{P}_t$ have been calculated and the cycle starting at Equation 3.11 can repeat.

The UKF is able to make predictions that closely fit the simulated experimental data based on the Lorenz system even when parameter values are unknown (Figure 24). The parameter values themselves are also accurately estimated in the process (Figure 25). These observations with the Lorenz system have been made before [118, 124]. However, there has been no mention that the promising results are dependent on carefully chosen values for the system noise $\boldsymbol{\varphi}$, the initial state $\mathbf{x}_0$ and the initial estimate of parameter values θ_0. In our experience,

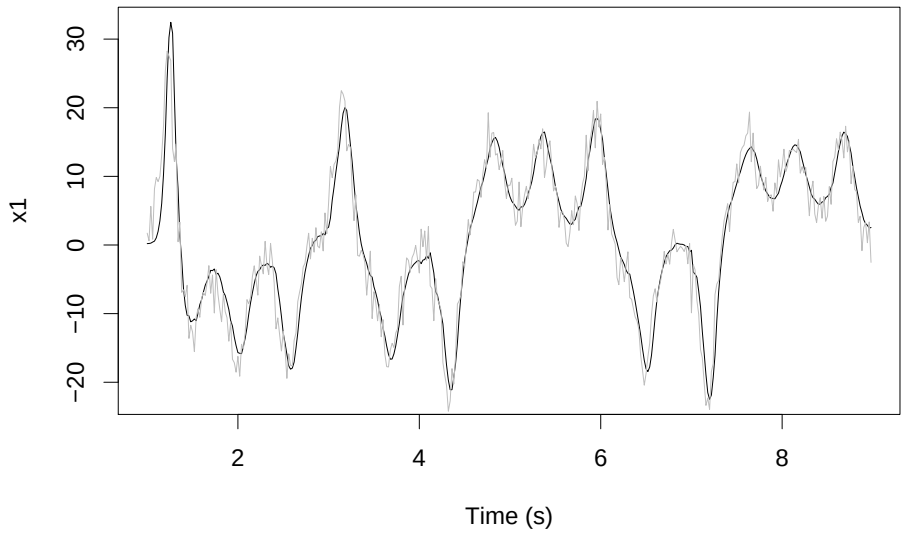


Figure 24: A fit of the Lorenz equations to the simulated experimental data using an Unscented Kalman Filter. The simulated experimental data is given in gray and the fitted model in black.

when initial conditions or system noise are not chosen appropriately the UKF may not converge when making time series predictions or may encounter numerical errors when taking matrix square roots.

The lack of robustness experienced with the unmodified UKF makes it unsuitable for use as an unsupervised parameter estimator on automatically produced mathematical models. However, the UKF can be made more robust by performing multiple runs of the filter with $\boldsymbol{\varphi}$, $\mathbf{x}_0$ and $\boldsymbol{\theta}_0$ under the control of an optimiser as shown in algorithm 5. We found that a particle swarm optimiser of the type described by He et al. [43], and shown in Algorithm 1, found usable UKF settings after 10 iterations using only 10 particles. Figures 24 and 25 were produced with UKF settings that had been calculated using a particle swarm optimiser.

Particle Filter

Particle filters offer an alternative to the Unscented Kalman filter that do not assume Gaussian distributions and can handle non-Gaussian errors such as the residuals produced by a model that does not completely describe the behaviour of experimental data. We found that particle filters were more robust than the UKF and did not require

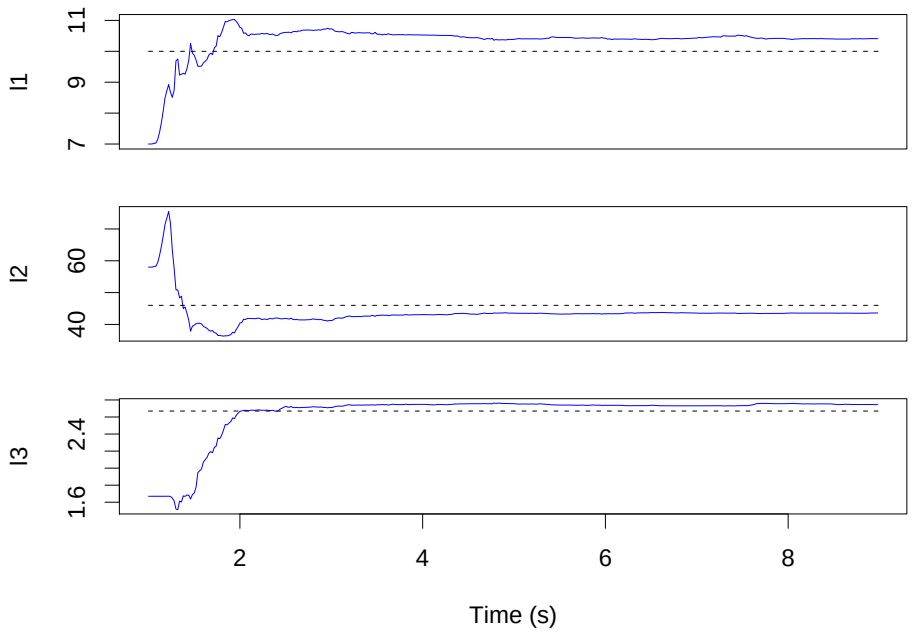


Figure 25: Predicted parameter values during a run of the Unscented Kalman Filter shown as blue lines ($l1 = \lambda_1$, $l2 = \lambda_2$, $l3 = \lambda_3$). The actual value of the parameters used to generate the simulated experimental data are given as black dotted lines.

Algorithm 5 Optimised Unscented Kalman Filter

Initialise minimum cost: C_b
Initialise best: $\boldsymbol{\varphi}_b, \mathbf{x}_{0b}, \theta_b$
for $i = 1 \rightarrow N_{runs}$ **do**
  Obtain $\boldsymbol{\varphi}$, $\mathbf{x}_0$ and θ_0 from optimiser.
  Run Unscented Kalman filter to get cost C
  if $C < C_b$ **then**
    $\boldsymbol{\varphi}_b, \mathbf{x}_{0b}, \theta_b \leftarrow \boldsymbol{\varphi}, \mathbf{x}_0, \theta_0$
  end if
  Update optimiser with C
end for
Run filter with $\boldsymbol{\varphi}_b, \mathbf{x}_{0b}, \theta b$
Return final filter estimate of parameters, θ_M

restarting with different setups to get the filter to closely track the data.

What follows is a description of the resample move particle filter [34] which was found to perform better than a conventional condensation particle filter [49] for parameter estimation.

Particle filters conventionally work with a time series which is assumed to be generated by the state space model described by equations 3.7 and 3.8. Since the equations we are working with are deterministic, we used a particle filter based on a simpler model which has no process noise:

$$\mathbf{x}_t = f(\mathbf{x}_{t-1}, \theta_j) \tag{3.30}$$

$$z_t = h(\mathbf{x}_t, \varepsilon)\,. \tag{3.31}$$

A conventional use for a particle filter is to estimate $p(\mathbf{x}_t|z_{0:t}, \theta)$ by importance sampling over a set of particles $\{\mathbf{x}_t^{(k)}\}_{k=1}^{N_p}$. Each particle contains a hypothesised state of the system i.e. the values of the variables. Particle filters can be made to estimate the parameters of a model, θ, using the particles $\{\mathbf{x}_t^{(k)}, \theta^{(k)}\}_{k=1}^{N_p}$ to get $p(\mathbf{x}_t, \theta|z_{0:t})$.

The resample move particle filter is described in Algorithm 6. It uses weighted particles to get an estimate of $E[\mathbf{x}_t]$ and $E[\theta]$. In order to keep the particles relevant, they are resampled on every iteration to remove particles with low weights.

An effect known as "sample impoverishment" occurs when particles cluster around a very small area of state space. In order to combat this, a move step is used by the particle filter where each particle is taken through a Monte Carlo Markov Chain (MCMC) transition in order to explore state space more fully. We used slice sampling as an MCMC step. Slice sampling doesn't rely on a carefully chosen control parameter, unlike some alternatives such as the Metroplis-Hastings algorithm, making it more convenient for automated parameter estimation.

Parameter estimates performed with a resample-move particle filter are not as stable or accurate as those obtained from an Unscented Kalman filter (Figure 26). However, the particle filter still tracks the input signal despite having imprecise values for the model parameters (Figure 28).

Algorithm 6 Resample-move particle filter

Require: randomly initialised particles $\{\mathbf{x}_0^{(k)}, \theta^{(k)}\}_{k=1}^{N_p}$
Require: uniform particle weights $\{\alpha^{(k)}\}_{k=1}^{N_p} \longleftarrow \frac{1}{N_p}$
1: **for** $t = 1 \rightarrow M$ **do**
2: **for** $k = 1 \rightarrow N_p$ **do**
3: Use the state space model to calculate $p(z_t|\mathbf{x}_{t-1}^{(k)}, \theta^{(k)})$
4: Perform an MCMC transition on $\mathbf{x}_{t-1}^{(k)}$ using $p(z_t|\mathbf{x}_{t-1}^{(k)}, \theta^{(k)})$
5: $\alpha^{(k)} \leftarrow \alpha^{(k)} \times p(z_t|\mathbf{x}_{t-1}^{(k)}, \theta^{(k)})$
6: $\mathbf{x}_t^{(k)} \leftarrow f(\mathbf{x}_{t-1}^{(k)}, \theta^{(k)})$
7: **end for**
8: $\alpha_\Sigma \leftarrow \sum_{k=1}^{N_p} \alpha^{(k)}$
9: Normalise weights: $\{\alpha^{(k)}\}_{k=1}^{N_p} \leftarrow \{\alpha^{(k)}\}_{k=1}^{N_p} \div \alpha_\Sigma$
10: Make a prediction: $E[\mathbf{x}_{t+1}] = \sum_{k=1}^{N_p} \alpha^{(k)} f(\mathbf{x}_t^{(k)}, \theta^{(k)})$
11: Replace particles that have small weights with higher weighted particles
12: **end for**

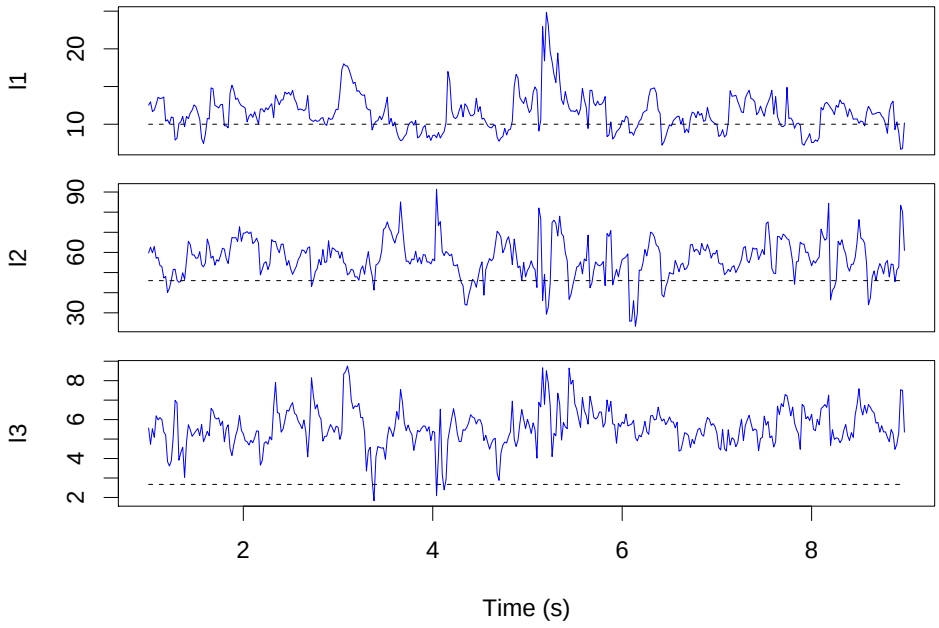


Figure 26: Predicted parameter values during a run of a resample-move particle filter shown as blue lines ($l1 = \lambda_1$, $l2 = \lambda_2$, $l3 = \lambda_3$). The actual value of the parameters used to generate the simulated experimental data are given as black dotted lines.

3.4 MODEL EVALUATION

3.4.1 *Evaluating Models by Prediction Performance*

A resample-move particle filter is able to make reasonable predictions of chaotic time series measurements even though it fails to make accurate estimates of the model parameters (Figure 28). A question that arises from this is whether the structure of the model, that is being fitted to the time series, is being exploited to make the predictions. If the model does have a significant effect, and the range of the parameter estimates made by the particle filter in Figure 26 suggests it does, then a particle filter could conceivably be used to evaluate a model by quantifying its prediction performance for some experimental data. Ideally, a good model should give accurate predictions and the most suitable model should give the best predictions.

To see if prediction performance could be used to evaluate models, we investigated whether the Lorenz equations score better than alternatives for the simulated experimental data shown in Figure 17. To select alternative equations, a genetic programming mutation operator [100] was used on the original Lorenz equations. Even heavily mutated equations scored as well as, or better than, the equations that were used to generate the time series (Figure 28). This behaviour makes particle filter prediction performance an unsuitable criterion for evaluating models.

3.4.2 *Evaluating Models after Fitting*

Algorithm 7 Fitting models using a UKF

Require: a training set of data $\{z\}_{t=0}^{M}$
Require: a testing set of data $\{w\}_{t=0}^{M}$
  estimate θ from $\{z\}_{t=0}^{M}$
  predict $\{y_{t+1}\}_{t=0}^{M-1}$ from $\{w\}_{t=0}^{M}$ using θ
  return $\sum_{t=1}^{M}(y_t - w_t)^2$

Of the parameter estimation methods considered, the Unscented Kalman Filter (UKF) described in section 3.3.2 estimates parameters the most accurately. A set of parameters can be estimated, using a UKF, from a training set of simulated experimental data. These parameter values are then fixed and used on testing sets of simulated experimental data to make predictions. The accuracy of the predictions obtained for the testing sets gives a score to how well a given model fits the

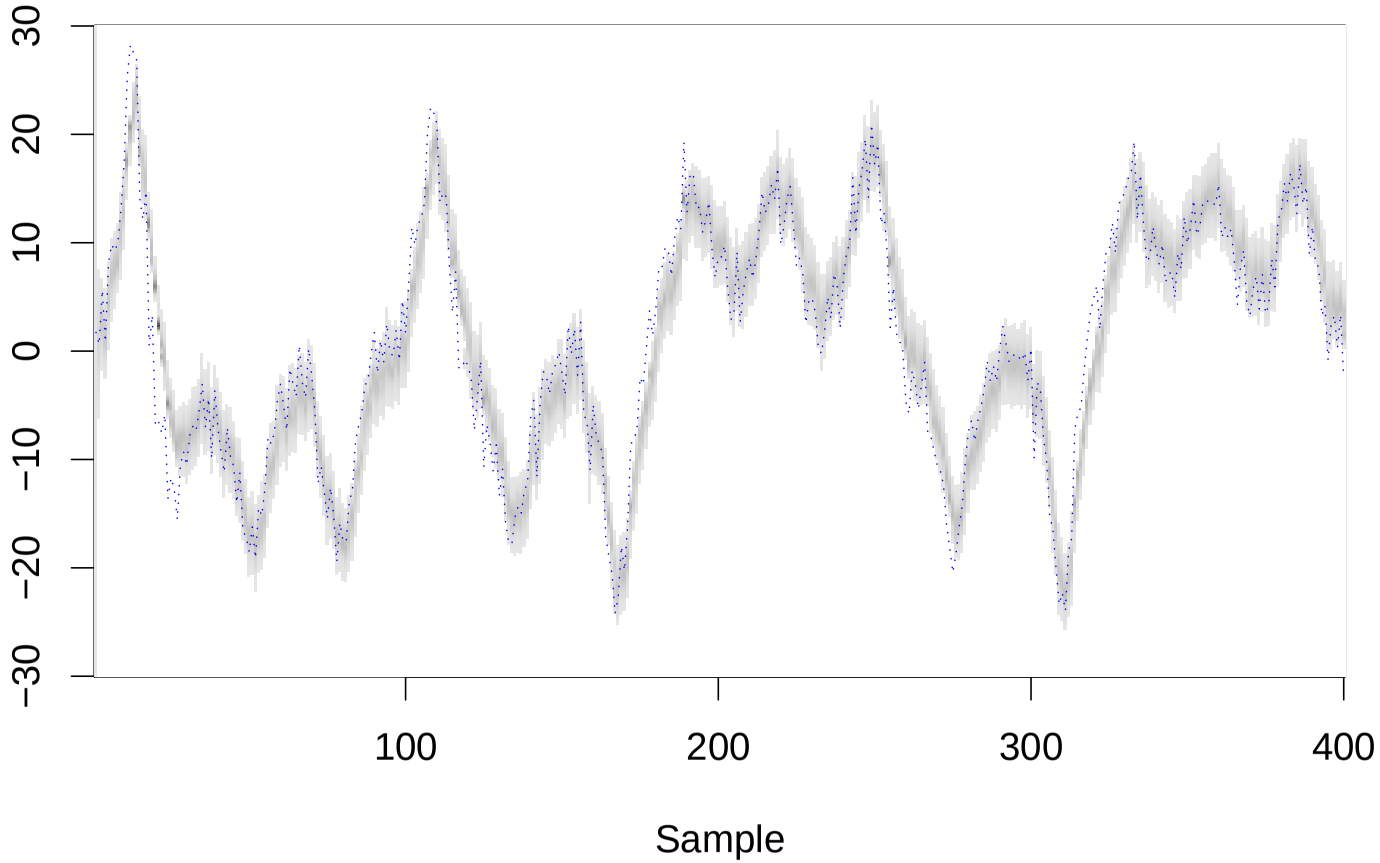


Figure 27: Probability density of y_{t+1} calculated by a particle filter when given z_t for simulated experimental data generated from the Lorenz equations. Darker colours indicate a higher probabilities. The probabilities were calculated over 200 particles. The blue dotted line indicates the actual value of the time series at z_{t+1}. The x-axis indicates the sample number with a sampling period of 0.02 seconds.

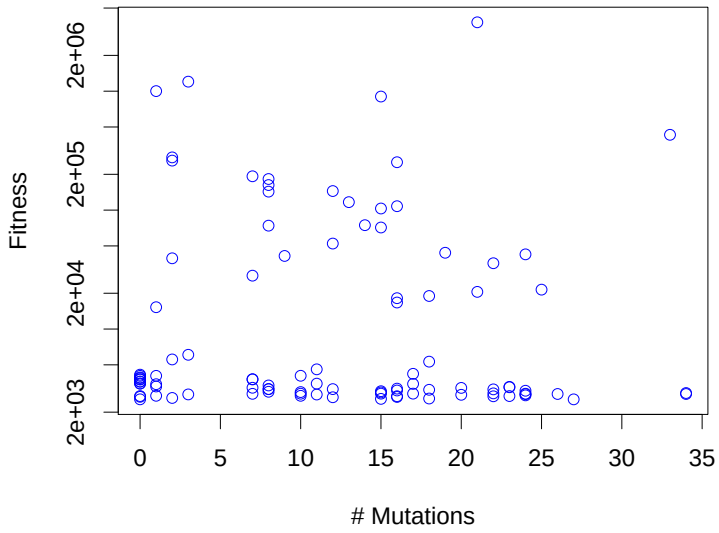


Figure 28: Fitness values for mutated models of the Lorenz equations obtained using the prediction performance of a resample-move particle filter. A lower fitness score indicates more accurate predictions and a more suitable model. The fitness score was calculated as $-\log(p(D|\mathcal{H}_j))$ where D is the simulated experimental data and $\{\mathcal{H}\}_{j=1}^{100}$ are the models being considered.

available data (Algorithm 7). This method differs from the particle filter method described in section 3.4.1 in that accurate parameter estimates are being made and that training (estimating parameters) and testing (scoring) are done over different time series. It is expected that this approach will reduce the possibility of overfitting — a potential cause of which could be parameter fluctuations.

The UKF fitting was tested as a model evaluation score using the same technique that was applied to the particle filter in section 3.4.1. Out of 180 mutants, 2 sets of equations performed better than the Lorenz set of equations on the simulated experimental data. The two false positives rule out the possibility of using the UKF to identify a definitive model from experimental data. However, this method could be used to find a small set of viable models from a larger set of candidate models.

3.5 MODEL GENERATION

Previous sections in this chapter described methods of evaluating how well a particular model fits a times series of experimental data. This section describes methods of generating and optimising multiple models to propose suitable candidates to explain the data.

3.5.1 *Genetic Programming*

Genetic Programming (GP) is a technique which breeds structures, representing equations or computer programs, and optimises them to reach a specified goal. Here we consider the use of GP to generate equations which are represented as trees (Figure 29).

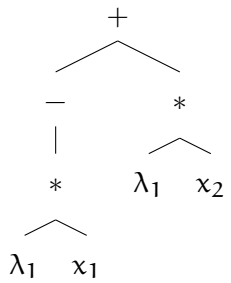


Figure 29: Equation 3.3, $\dot{x}_1 = -\lambda_1 x_1 + \lambda_1 x_2$, represented as a tree for use in genetic programming

Usually, GP works on a population of trees and selectively applies the operations of mutation and crossover to the population based on the results of evaluating a fitness function. The mutation operator takes part of a tree and replaces it with a new subtree recognising the constraint that the result must be a valid tree. The crossover operator takes two existing trees and combines them to create new trees. Equations that have a better fitness have a higher probability of being chosen for crossover.

Many variations exist for the different algorithms that make up a GP framework. Here, we keep to the widely used algorithms shown in Table 3.

Exploring Model Space

The production and evaluation of potential models can be thought of as an exploration of a model space. Unlike the parameter space described in section 3.3.1, this model space does not have obviously quantifiable dimensions. In order to gauge the difficulty of exploring

Algorithm 8 Edit distance between trees a and b — $e(a, b)$

$t(x)$ is the total number of nodes in tree x
if $t(a) = 0$ **then**
  Return $t(b)$
end if
if $t(b) = 0$ **then**
  Return $t(a)$
end if
if a and b are identical **then**
  Return 0
end if
$m \leftarrow$ number of elements in top level of a
$n \leftarrow$ number of elements in top level of b
create matrix $M_{m+1 \times n+1} = 0$
for $i = 1 \rightarrow m$ **do**
  $M_{i,0} \leftarrow M_{i-1,0} + t(a_{i-1})$
end for
for $i = 1 \rightarrow n$ **do**
  $M_{0,i} \leftarrow M_{0,i-1} + t(b_{i-1})$
end for
for $i = 1 \rightarrow m$ **do**
  for $j = 1 \rightarrow n$ **do**
    set the delete cost $D \leftarrow t(a_{i-1})$
    set the insert cost $I \leftarrow t(b_{j-1})$
    set the substitution cost $S \leftarrow e(a_{i-1}, b_{j-1})$
    $M_{i,j} \leftarrow \min\left(M_{i-1,j} + D, M_{i,j-1} + I, M_{i-1,j-1} + S\right)$
  end for
end for
Return $M_{m,n}$

Table 3: Algorithms used to perform GP on the Lorenz equations.

Algorithm	Description
Ramped half-and-half	Initialise half the population with full trees, that have all leaves at a maximum depth, and half the population with trees that have been grown to have more asymmetrical shapes.
Tournament Selection	Select x individuals from the population at random. The single fittest of those individuals takes part in crossover.
Subtree Crossover	Select a crossover point in each parent tree and swap the subtrees rooted at the crossover point.
Point Mutation	Mutate a randomly selected point in a tree with a randomly generated subtree.

the model space with GP, a test was performed with a mock fitness function.

The mock fitness function calculates the edit distance between each differential equation in a generated model and the Lorenz system of equations. The algorithm used to calculate the edit distance between 2 single equations is given in Algorithm 8. This fitness is cheap to calculate and is a well behaved in the sense that models which have a similar structure to the Lorenz equations will have a small edit distance and will therefore have a higher probability of being selected by the GP algorithms. The goal of the GP system using the mock fitness function is recovery of the Lorenz equations.

GP was performed 50 times with the mock fitness function to quantify the effects of population and the number of generations on reaching the goal of recovering the Lorenz equations (Figure 30). The number of generations did not impact the quality of the final result. For instance, the results after 500 generations are no better than the ones after 20 generations. The results are dependent on population size and improve until the population exceeds 10000 equations. However, even large populations do not guarantee that the original equations will be recovered.

In the unlikely event that a fitness function based on parameter estimation and model fitting will behave as consistently as the mock fitness function, the results suggest that a population of at least 10000 will be needed over 20 generations to recreate a set of equations from simulated test data. This large population would result in a

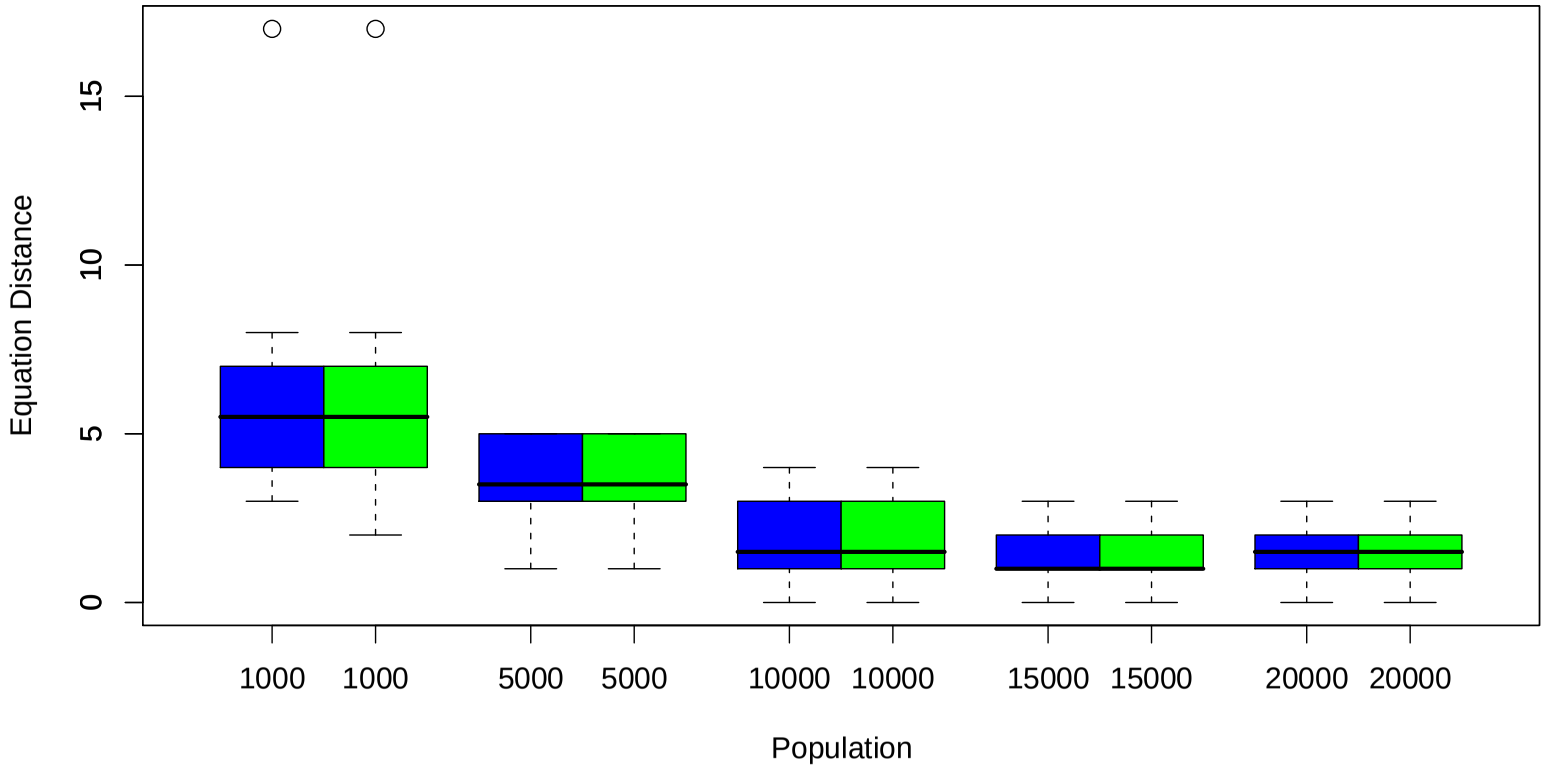


Figure 30: Box plot of the final equation distance after 50 runs of genetic programming using a mock fitness function. The population used is shown on the x-axis. Boxes are drawn from the first to the third quartiles and the medians are shown as heavy lines. Dotted whiskers indicate the range of the data and circles mark outliers. The results after 20 generations are shown in blue and 500 generations are shown in green.

total of 200000 evaluations of a fitness function. Taking an estimate of 10 CPU minutes to calculate the parameters for a single set of chaotic equations, the recovery of the Lorenz system from a time series would take 198 CPU weeks. The large CPU requirement of even this optimistic estimate would require more time than is currently available to UK researchers through the national computing grid. As suggested in Figure 30 such a large undertaking may not give the correct answer in the majority of cases.

3.5.2 *Inductive Process Modelling*

The numerical experiment performed in section 3.5.1 illustrates the difficulty in exploring a large model space. It is possible to reduce the model space by constraining the GP system to only consider equations with the correct units of measurement [59, 5]. Another way, that greatly reduces model space, is to only examine models that are regarded viable by an expert on the system being considered.

Inductive process modelling (IPM) [70, 14] is a machine learning technique which combines parts of an ODE system, or processes, into alternative sets of equations. These induced equations can then be ranked based on their fit to experimental data. For example, when analysing Ca^{2+} oscillations in animal systems the following processes could be considered [113]:

- The leak of Ca^{2+} into the cell.
- The pumping of Ca^{2+} out of the cell.
- The release of Ca^{2+} from the ER into the cytosol.
- The transport of Ca^{2+} into the ER by sarco-/endoplasmic reticulum Ca^{2+} ATPase.
- The buffering of Ca^{2+} by cytosolic proteins.
- The sensitivity of IP3R channels to Ca^{2+} induced Ca^{2+} release.

Each of these processes could be modelled with one or more alternative sets of equations. For instance the Ca^{2+} ATPases could be modelled using a Hill function or simplified to a linear term. Some processes could be optional in a model, for example few ODE models of Ca^{2+} oscillations take into account the buffering of Ca^{2+} to cytosolic proteins.

An inductive process modelling system combines permutations of these processes to produce a space of models that is typically small

enough for each model to be assessed. An example of this is given by Bridewell et al. [14] for the Ross Sea ecosystem where processes are combined to produce a total of 1024 models. Despite being a far more complex model structure than the 3 parameter Lorenz equations, this number of model evaluations is a tenth of that produced in a single generation of genetic programming as considered in section 3.5.1.

We evaluated IPM on a non-chaotic Ca^{2+} oscillating system. The One Pool model [21] was used to generate a noiseless time series for cytosolic Ca^{2+} concentration. Using the processes and alternative equations given in Table 1 of the review by Schuster et al. [113] we specified background knowledge about the system.

For example, V_{in} the rate of influx of Ca^{2+} into the cytosol could be modelled as a constant flux across the plasma membrane, ν_0, plus an IP3 mediated release, $\nu_1\beta$. As an alternative V_{in} could be set to 0:

$$V_{in} = [\nu_0 + \nu_1\beta] \bigvee 0 \tag{3.32}$$

The $\bigvee$ denotes that V_{in} could be set to either of the two terms.

Using the same notation, the alternatives for other rate laws from [113] can be stated:

$$\begin{aligned}
V_{out} &= kC_{cyt} && (3.33)\\
V_{rel} &= \left[k_f C_{er} + \beta V_{M3}\frac{C_{er}^2}{K_R^2 + C_{er}^2}\frac{C_{cyt}^4}{K_A^4 + C_{cyt}^4}\right] \\
&\quad \bigvee \left[\left(k_0 + k_1 R\left(\frac{C_{cyt}^2}{K_1^2 + C_{cyt}^2}\right)^3\right)(C_{er} - C_{cyt})\right] \\
&\quad \bigvee \left[\left(k_{leak} + k_{ch}\frac{C_{cyt}^2}{K_1^2 + C_{cyt}^2}\right)(C_{er} - C_{cyt})\right] && (3.34)\\
V_{serca} &= \left[V_{M2}\frac{C_{cyt}^2}{K_2^2 + C_{cyt}^2}\right] \bigvee [k_{pump} C_{cyt}] && (3.35)\\
V_{rec} &= k_3(1 - R) && (3.36)\\
V_{des} &= k_{-3} C_{cyt} R && (3.37)\\
V_b &= k_+(B_0 - B)C_{cyt} - k_- B && (3.38)
\end{aligned}$$

Similarly, the alternative differential equations for Ca^{2+} oscillations are:

$$\begin{aligned}
\frac{dC_{cyt}}{dt} &= [V_{in} - V_{out} + V_{rel} - V_{serca} - V_b] \\
&\quad \bigvee [V_{in} - V_{out} + V_{rel} - V_{serca}] && (3.39)\\
\frac{dC_{er}}{dt} &= \rho_{er}(V_{serca} - V_{rel}) && (3.40)
\end{aligned}$$

$$\left[\frac{dB}{dt} = V_b\right] \bigvee \emptyset \tag{3.41}$$

$$\left[\frac{dR}{dt} = V_{rec} - V_{des}\right] \bigvee \emptyset \tag{3.42}$$

Where $\emptyset$ denotes that, for some models, the differential equation is not present and the variable it describes is not in the system.

These alternative sets of equations generated 96 models in total. Unviable models, which used variables which did not have a differential equation term, were removed. Redundant models, which contained a differential equation for a variable that wasn't used elsewhere, were also removed. This left 24 possible models, three of which are described in literature [21, 73, 77].

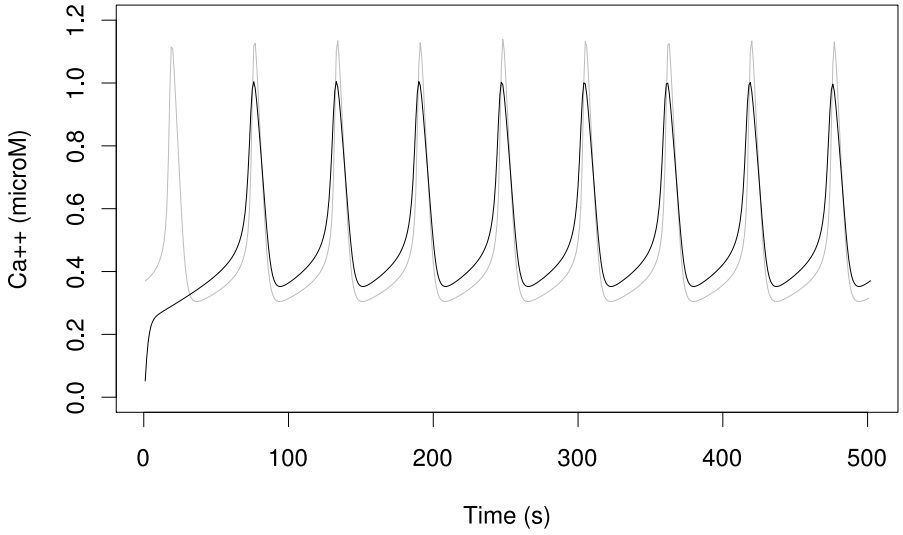


Figure 31: A fitted model (black) to noiseless data generated by the one pool model (gray). Hypotheses were generated by inductive process modelling using [113] as domain knowledge. Each potential model was fitted to the data using single shooting with a particle swarm optimiser. The fit shown is for the best scoring model which contains equations identical to the One Pool Model.

Each model was fitted to the simulated experimental data using the single shooting method with a particle swarm optimiser containing 10000 particles run over 20 generations. Models were evaluated using the Akaike information criterion (AIC) [1] that takes into account the number of parameters in each model as well as the square of the residuals of the fit. Of the 24 models evaluated, the best scoring

model was the One Pool model (Figure 31). This successful recovery of the equations is possible given the small number of candidate models produced by IPM and the relative ease of fitting to noiseless non-chaotic data.

3.6 SCALING UP TO A LARGER MODEL

The numerical experiments run in section 3.4.2 and section 3.5.2 suggest that it is possible to identify a small group of models that could be used to explain a time series of chaotic data. However, the only chaotic system considered has been the Lorenz set of equations which requires only 3 parameters to be fitted. Existing chaotic models of intracellular Ca^{2+} oscillations have more than 10 parameters and so it is necessary to investigate if the technique described in section 3.4.2 can be scaled up to a more complex system of equations.

We considered the chaotic Ca^{2+} model proposed by Haberichter et al. [39] containing 3 variables and with 14 parameters set to unknown values. Simulated experimental data was generated by integrating the model with 20% additive noise. The original equations were mutated 100 times and the fitness calculated using the sum of squares between the UKF predictions and a test set of simulated experimental data. A scatter plot of the results is shown in Figure 32. A total of 12 mutants had a better fitness than the original equations suggesting that parameter fitting and model evaluation is problematic with realistic chaotic models of Ca^{2+} spiking.

3.7 DISCUSSION

This chapter describes the investigation into whether automated methods could be used to suggest viable models for a chaotic Ca^{2+} spiking system using two general procedures shown in Figure 16 — model proposal and model fitting. Even though fewer techniques have been developed for model proposal, this subproblem was the most surmountable and Section 3.5.2 demonstrated the successful recovery of the equations describing a periodic spiking system. Unfortunately, the other problem of fitting a set of equations to chaotic data was much harder to solve. Only a simple chaotic model with few parameters could be fit to simulated experimental data well enough for the equations to be identified.

The shortcomings in parameter estimation suggest that any further research into chaotic systems identification with regards to the Ca^{2+}

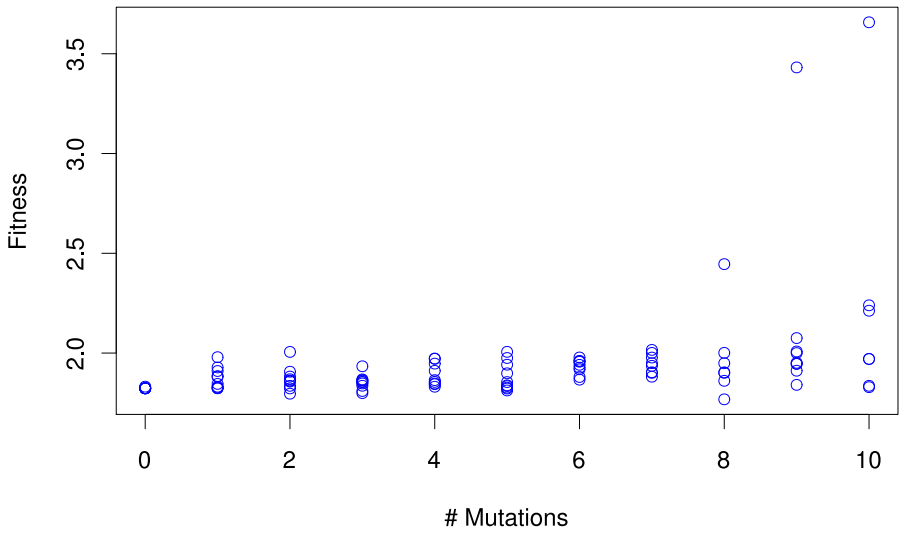


Figure 32: Fitness values for mutated models of the Haberichter chaotic Ca^{2+} spiking model obtained using the prediction performance of a Unscented Kalman filter. A lower fitness score indicates more accurate predictions and a more suitable model. The fitness score was calculated as a sum of the squares of the residuals.

spiking data should concentrate on improving performance on models with over 10 parameters. There is plenty of scope for continuing the investigation as the parameter estimation algorithms considered are not comprehensive. One notable omission is an investigation into using the multiple shooting algorithm [4]. This algorithm has been successfully used on the Lorenz system of equations but no publically available implementations exist. The optimisation algorithm is complex, in either its deterministic [12] or Bayesian forms [85]. Building a multiple shooting implementation would be time consuming with no certainty of success on a chaotic system with a larger number of parameters. A greatly simplified version of the multiple shooting algorithm was developed during this investigation that used a hybrid optimiser [106] consisting of an evolutionary strategy [108] and a nonlinear interior point method [89]. This simpler multiple shooting algorithm did not estimate parameters as accurately as the UKF (Table 4). A fit produced by the simplified multiple shooting algorithm to the Lorenz equations is shown in Figure 33.

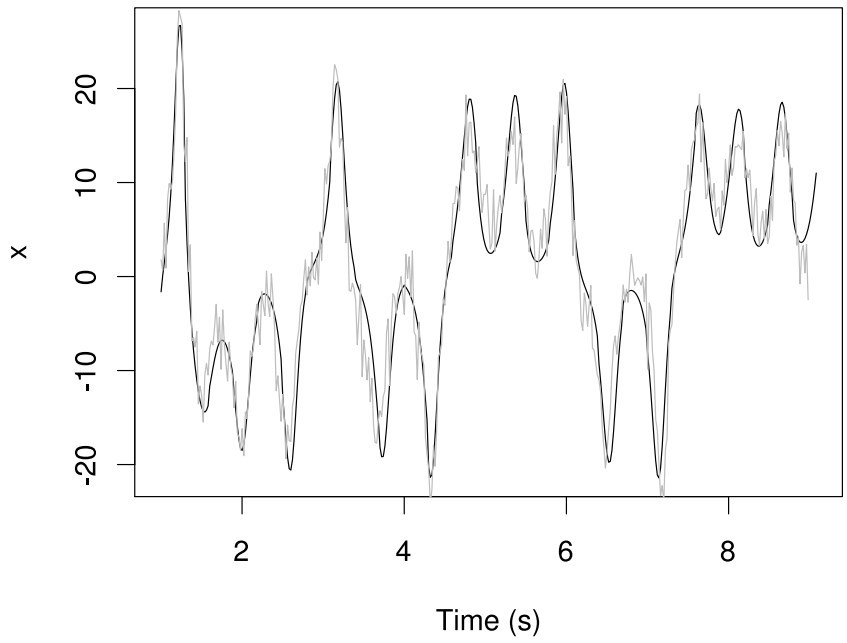


Figure 33: Fit (black) to simulated experimental data (gray) using a multiple shooting algorithm. The algorithm had a segment size of 15 sample points and used an SRES optimiser for 100 generations followed by a nonlinear interior point optimisation to find a local minima.

Another candidate technique is the ensemble Kalman filter [27] which has been successfully used to estimate the state of real chaotic systems such as hurricane vortices [19]. However, using this technique on the Lorenz system of equations proved to be time consuming with parameter estimation taking around 24 hours with an ensemble size of 10000. The parameter estimates made by the ensemble Kalman filter were less accurate than those produced by the UKF (Table 4).

Table 4: The mean normalised standard error (MNSE) of algorithms used for parameter estimation. The normalised standard error for each parameter estimate is calculated as $\sqrt{\frac{(e-a)^2}{a^2}}$ where e is the estimated parameter value and a the actual parameter value.

Algorithm	Optimisers	MNSE
PF (Condensation) [49]	—	Fail
Single Shooting	SRES [108]	0.60
PF (Gilks) [34]	—	0.51
Single Shooting	Particle Swarm [43]	0.41
Multiple Shooting [4]	SRES[108]/interior point[89]	0.34
Ensemble KF [27]	—	0.31
Unscented KF [55]	—	0.05

Because the investigation failed to distinguish a suitable parameter estimation algorithm, there was no comparison of methods that could be used to score a potential model. Simply using the least squares fit of a model to measure its suitability to describe given data does not take into account model parsimony. Because more complex models are able to fit a larger variety of data, scoring with the square of the residuals can result in large and unrealistic models being identified and also leads to the possibility of overfitting. Several techniques exist that are able to satisfactorily capture the trade-off between model complexity and accuracy when scoring a model. A widely used example is the Akaike Information Criterion used in Section 3.5.2. However, alternatives such as the Bayesian Information Criteria [114] and the Minimum Description Length [40] could also be used as a scoring scheme.

Genetic Programming (GP) didn't appear a suitable technique for model generation as demonstrated in Section 3.5.1. However, GP is a wide area of research and the investigation described here only analysed one combination of widely used GP ingredients. The most significant problem with GP was that it generated a large proportion

of unviable models resulting in an inefficient use of the expensive model evaluation step. Some variations of GP address this problem by only generating equations with the correct dimensions [94], or speeding up model evaluation in young populations of equations [6]. These, more efficient, techniques were not investigated because the published improvements in efficiency were insignificant compared to the performance of Inductive Process Modelling (IPM). The success of IPM in identifying a periodic model, with an almost optimal number of model evaluations, suggest it as is a very promising model generation technique despite its meager use in the field of systems biology.

3.8 SOFTWARE USED

A program written in C++ was used to generate the data for plots of parameter space and parameter estimates. This program used the LSODA [99] implementation provided by Heng Li at the Wellcome Trust Sanger Institute to integrate time series. The Unscented Kalman filter and condensation particle filter implementations from the DYSII Dynamic Systems Library were provided by L.M Murray at CMIS, CSIRO in Perth, and the Stochastic Ranking Evolutionary Strategy implemented in libSRES [52] was used for parameter estimation. A particle swarm optimizer, a resample move particle filter and multiple shooting were implemented from descriptions given in literature. The multiple shooting parameter estimator used a hybrid optimiser that included a nonlinear interior point optimiser implemented by the OPT++ library provided by J.C Meza at Sandia National Laboratories. An overview of the parameter estimation software is given in the Appendices (Chapter 3).

The investigation into Genetic programming was implemented in ANSI Common Lisp running on the Steel Bank Common Lisp environment. Inductive Process Modelling was implemented as a python program that generated C++ libraries for use by the parameter estimation program described above. The investigation into Model mutation was also implemented in a python program that generated C++ libraries for use with the parameter estimation methods.

4

MODELS OF CALCIUM SPIKING

4.1 OVERVIEW

This chapter describes a periodic model for the Ca^{2+} spiking consisting of two differential equations. In order to produce the model, assumptions have to be made about unknown components. At the end of the chapter these assumptions and alternative possibilities are discussed.

4.2 INTRODUCTION

Chapter 3 discussed possible methods for automatically generating a model from chaotic experimental data. However, in this chapter a more conventional modelling approach is used. As discussed in Section 1.2.4, not much is known about the components that make up the oscillating system. However, by making parsimonious assumptions, this approach produces a good fit to a single Ca^{2+} spike (Figure 37) and suggests a mechanism in which the cation channel, DMI1, is essential for Ca^{2+} spiking.

We start by modelling the Ca^{2+} oscillations within the nucleus of plant root hair cells during symbiosis. The nuclear envelope is assumed to be a store of Ca^{2+} with a high concentration difference forcing Ca^{2+} out of the envelope and into the nucleus when Ca^{2+} channels open (Figure 40). The model is mostly electrical in nature and membrane potential plays a significant role in the modelled Ca^{2+} oscillations.

Mathematically the proposed model is similar to a basic model of oscillating action potential in a pancreatic β cell [9] which has been relocated to the nucleus of a plant root hair cell. In the model discussed here, the role of the K^+channel is performed by DMI1 which balances membrane potential so that Ca^{2+} can be released from the nuclear envelope.

The model contains a hypothesised voltage gated Ca^{2+} channel. Evidence for the existence of such channels has been found in plant nuclei [38] but they have not been linked to symbiotic Ca^{2+} spiking. To keep the model simple and specific, DMI1 is modelled as a Ca^{2+}

activated K^+ channel rather than a cation channel with unknown activator. This decision is consistent with homology modelling of a pea ortholog of DMI1 that suggested possible K^+ selectivity in the channel region and potential Ca^{2+} binding pockets in a gating ring region of the protein complex [23].

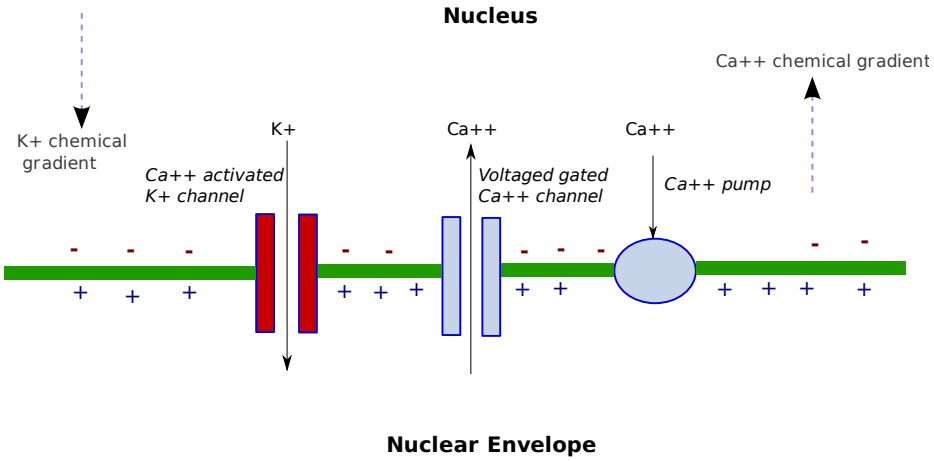


Figure 34: The components described by the simple model of nuclear Ca^{2+} spiking.

4.3 SIMPLE MODEL

In the model equations, lowercase letters are variables and parameters are denoted by uppercase or Greek letters. Ion channels are assumed to be ohmic, $i = gv_d$, where i is the current through the channel, g is the conductivity of the channel and v_d is the potential difference across the channel. Conductances are given for the whole nucleus and not per unit area.

The components of the model can be seen in Figure 34 and an electrical view of the model is given in Figure 35.

The model consists of two ordinary differential equations that capture the behaviour of a Ca^{2+} channel, a K^+ channel and a Ca^{2+} pump on the nuclear envelope. The change in the voltage across the nuclear membrane, v, is described by,

$$\frac{dv}{dt} = \frac{1}{C_m}(i_1 - i_2)\,. \tag{4.1}$$

The change in concentration of free Ca^{2+} within the nucleus, c, can be written as,

$$\frac{dc}{dt} = E_{ps}(\alpha i_1 - \mu c)\,, \tag{4.2}$$

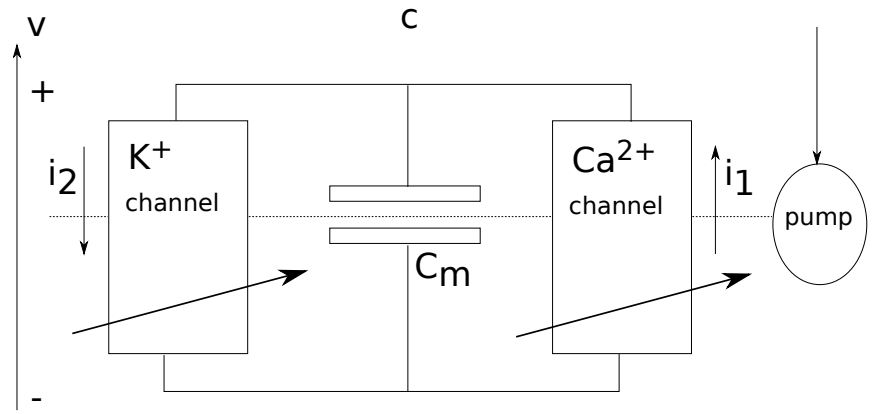


Figure 35: The simple model viewed as an electrical circuit. The dotted line indicates the location of the membrane of the nuclear envelope.

where i_1 is the Ca^{2+} current through the voltage gated Ca^{2+} channel and i_2 is the K^+ current through the Ca^{2+} activated K^+ channel — both with units fA. C_m is the capacitance of the nuclear envelope, E_{ps} is the proportion of free Ca^{2+} to buffered Ca^{2+} in the nucleus, μ is a pump rate and α is a value to convert from Ca^{2+} current to Ca^{2+} flux.

In a first approximation, the active transport of Ca^{2+} into the nuclear envelope is assumed to be electroneutral and does not directly contribute to membrane potential. A possible mechanism could be the countertransport of cations from the envelope into the nucleus so there is not net transport of charge.

The voltage gated Ca^{2+} channel is a hypothesised component of the system. The channel has a normalised voltage dependent conductance, $f(v)$, described by a Hodgkin-Huxley gate model [50]. There is an assumption of two activation gates per channel and the activation of the channel has a voltage dependence approximated by (Figure 36):

$$f(v) = \left(\frac{1}{1 + \exp\left(-\frac{v - V_{ml}}{K_{ml}}\right)} \right)^2 . \tag{4.3}$$

The current through the channel is dependent on the potential difference across the membrane and the resting voltage produced by the higher concentration of Ca^{2+} within the nuclear envelope,

$$i_1 = G_c f(v)(E_{ca} - v) . \tag{4.4}$$

The Ca^{2+} activated K^+ channel is assumed to cooperatively bind Ca^{2+} with two Ca^{2+} binding sites. The conductivity of the K^+ channel is described by a Hill function,

$$i_2 = G_k \frac{c^2}{c^2 + K^2}(E_k + v) . \tag{4.5}$$

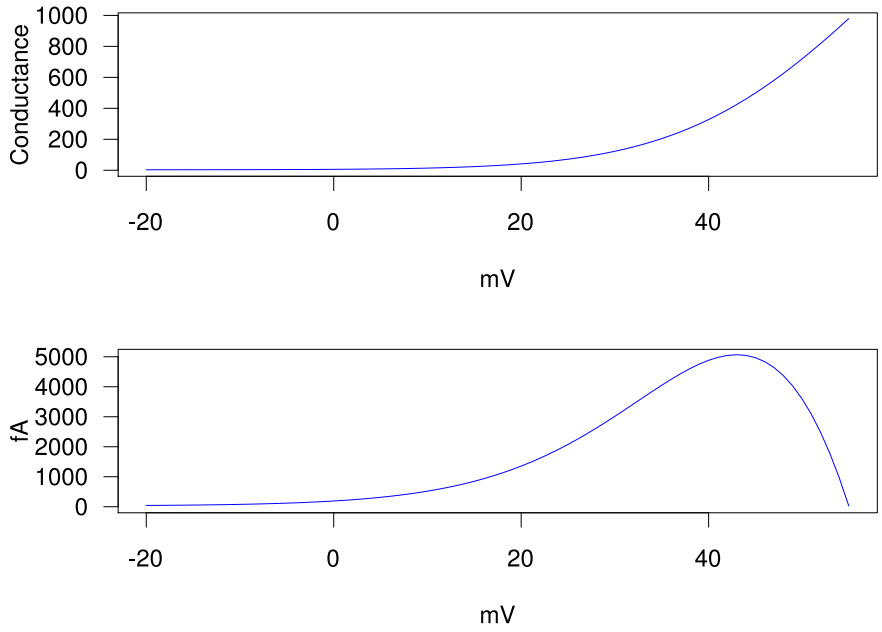


Figure 36: Voltage versus conductance (in pS) of the voltage gated calcium channel and the resulting Ca^{2+} current that flows through the channel assuming a 55 mV resting potential for Ca^{2+}.

4.3.1 *Parameters*

Many parameter values in the model are unknown. We therefore estimated some parameters by fitting the model to the Ca^{2+} spike shown in Figure 37. A single shooting algorithm using a particle swarm optimiser with 10000 particles was used. The parameter values obtained are given in Table 5.

4.3.2 *Overview of a Spike*

The model spontaneously oscillates through a range of biologically relevant initial conditions ($c = 40\,\text{nM} \rightarrow 1\,\mu\text{M}$, $v = -50\,\text{mV} \rightarrow 50\,\text{mV}$) indicating that it describes a fully activated system. To analyse a single spike we consider the initial conditions $c = 0.23\,\mu\text{M}$ and $v = -27\,\text{mV}$.

A phase portrait of the system, Figure 38a, shows that the chosen initial conditions lie at a lower voltage than the v nullcline (shown in blue). At $t = 0\,\text{s} \rightarrow t = 1.5\,\text{s}$ there is an rapid increase in the voltage across the membrane (v) brought about by a small Ca^{2+} leak current through the voltage gated channel (i_1). Despite the low conductance of the voltage gated channel, i_1 is driven by the combination of voltage

Table 5: Parameter values used in the simple model with their sources. The source 'Fit' indicates the parameter value was obtained by fitting the model to a Ca^{2+} spike

	Description	Value	Units	Source
V_n	Volume of nucleus	160	μm^3	[15]
C_m	Capacitance of nuclear envelope	5.1	pF	[38]
F	Faraday constant	10^{14}	$fC \cdot \mu mol^{-1}$	
α	Conversion of Ca^{2+} current to Ca^{2+} flux	0.03239	$\mu M \cdot fC^{-1}$	$\frac{1}{2FV_n}$
E_{ps}	Scaling factor relating total Ca^{2+} changes to changes in free Ca^{2+}	0.001		[15]
G_c	Total max conductance of voltage gated Ca^{2+} channels	2864	pS	Fit
V_{ml}	Half maximal activation of voltage gated Ca^{2+} channel	50.0	mV	Fit
K_{ml}	Constant in scaling function for voltage gated Ca^{2+} channel	14.7	mV	Fit
G_k	Total max conductance of Ca^{2+} activated K^+ channels	302	pS	Fit
K	Constant in hill function for Ca^{2+} activated K^+ channel	0.953	μM	Fit
μ	Pump rate into the nuclear envelope	24.9	s^{-1}	Fit
E_{ca}	Resting potential of Ca^{2+}	55	mV	[98]
E_k	Resting potential of K^+	17.7	mV	Fit

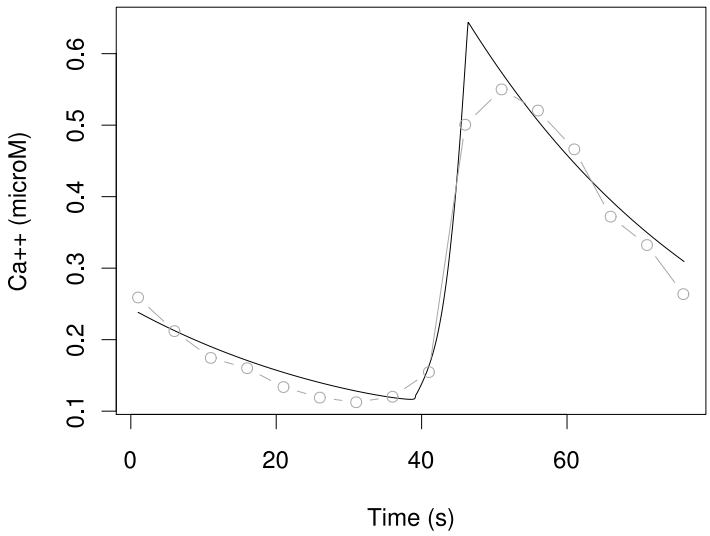
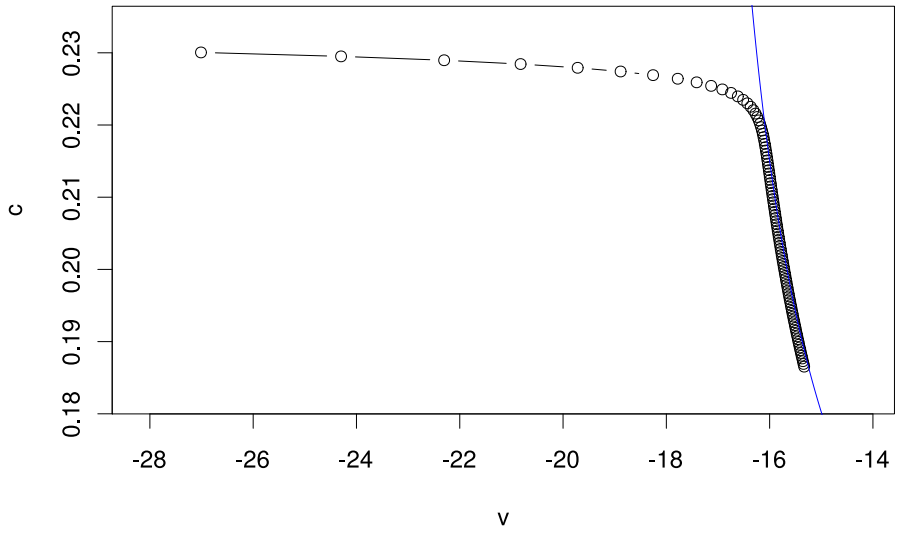


Figure 37: The simple model (black) fitted to a time series of a single Nod factor induced Ca^{2+} spike (gray).

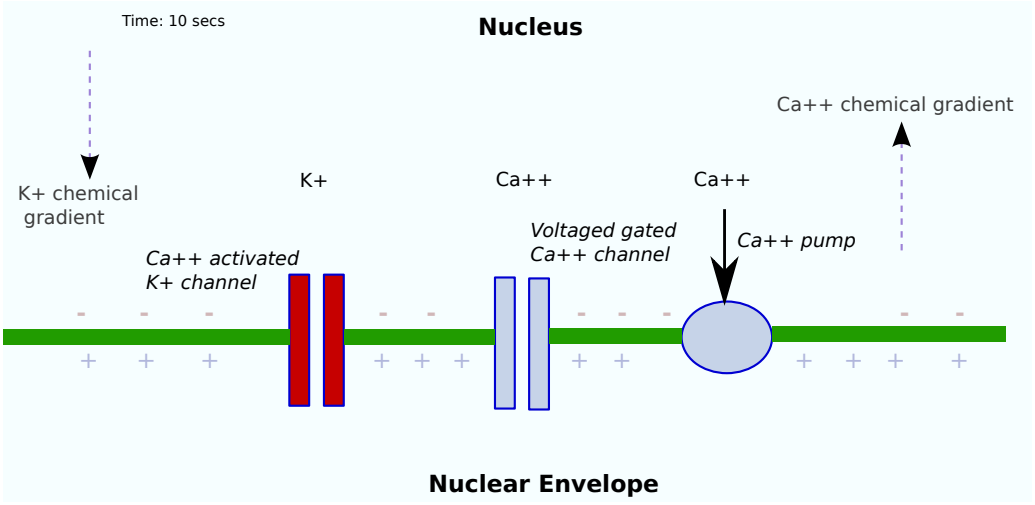
across the membrane and the concentration gradient of Ca^{2+}. When $t = 1.5\,s \rightarrow t = 10\,s$ the trajectory of the system closely follows the v nullcline with v still increasing but at a slower rate. On the v nullcline the membrane voltage doesn't change as the effects i_1 and the K^+ current (i_2) balance. However, in this region of phase space the Ca^{2+} in the nucleus (c) is decreasing as more Ca^{2+} is being pumped than is leaking through the voltage gated channel.

After some time the trajectory of the system diverges from the v nullcline and v begins to increase at a faster rate, Figure 39a. At $t \approx 31\,s$ the trajectory crosses the c nullcline shown in green. Crossing the c nullcline indicates that more Ca^{2+} is released by the voltage gated channel than is being pumped out of the nucleus ($\alpha i_1 > \mu c$, Figure 39b). This region of phase space lies at the start of a Ca^{2+} spike and is an area of positive feedback. The flow of i_1 raises v which increases the conductance of the voltage gated channel. This increased conductance results in a greater i_1. Because the K^+ channel is Ca^{2+} activated and c is increasing relatively slowly, i_2 is not large enough to balance the membrane voltage changes due to i_1.

An increasing v only results in a larger i_1 until $v \approx 40\,mV$ (Figure 36). If $v > 40\,mV$ the membrane voltage becomes significant enough to overwhelm any increased conductance of the voltage gated channel. At $v = 55\,mV$ the effects due to concentration gradient and membrane

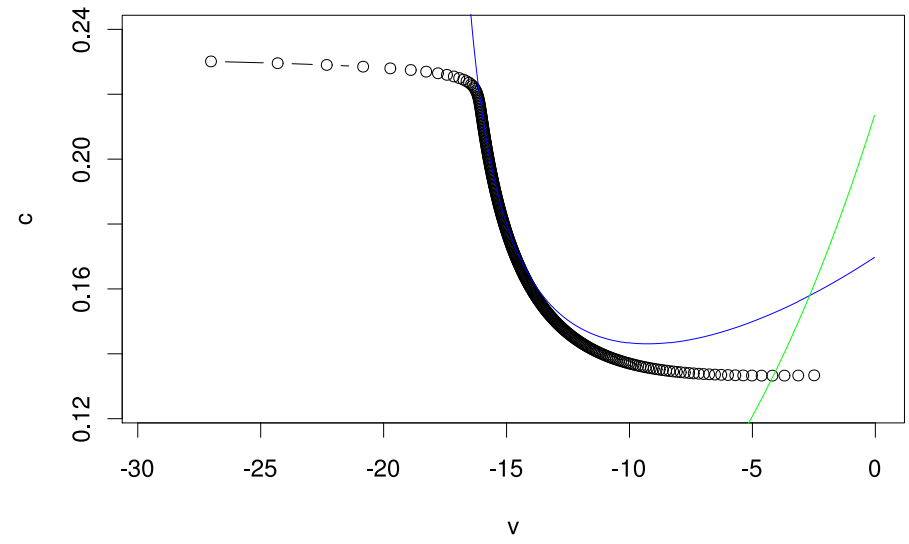


(a)

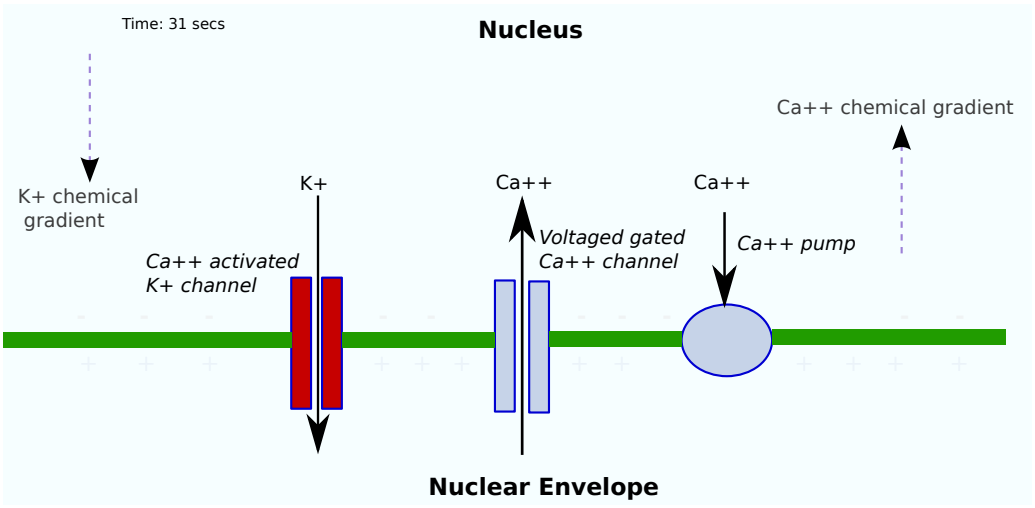


(b)

Figure 38: The simple model at $t = 10\,s$. (a) Trajectory of the model through phase space with the v nullcline shown in blue. Circles are plotted every 0.1 seconds. (b) Cartoon of the system showing Ca^{2+} and K^{+} currents as arrows with line width proportional to log of the current.

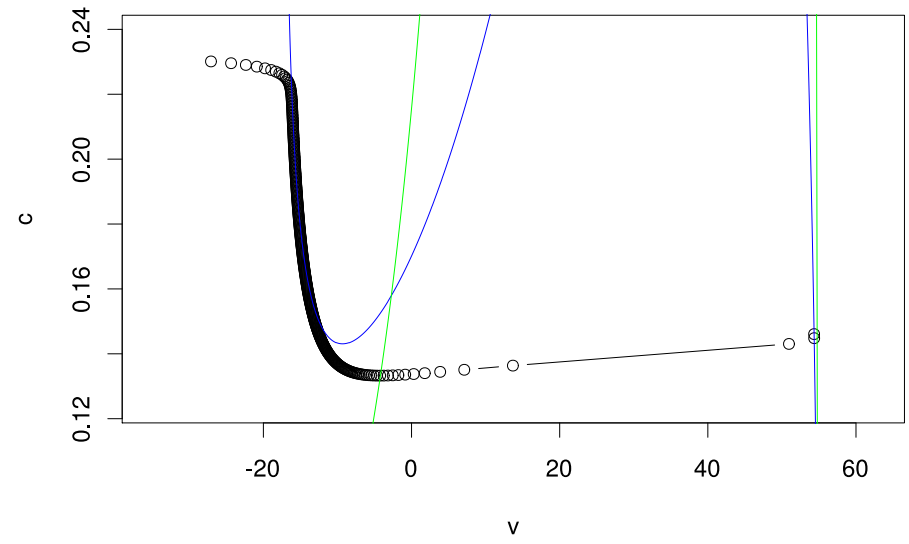


(a)

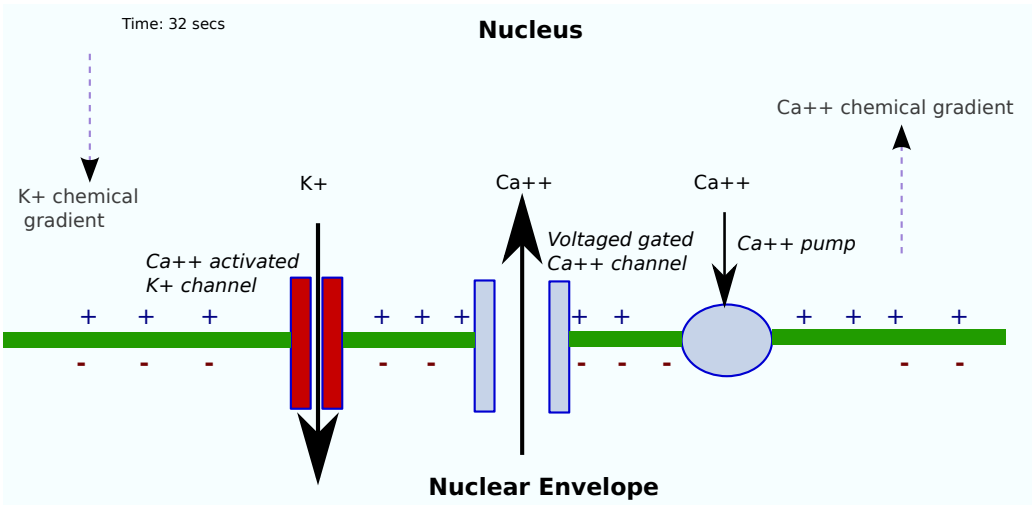


(b)

Figure 39: The simple model at t = 31 s . (a) Trajectory of the model through phase space with the v nullcline shown in blue and the Ca^{2+} nullcline shown in green. Circles are plotted every 0.1 seconds. (b) Cartoon of the system showing Ca^{2+} and K^{+} currents as arrows with line width proportional to log of the current.



(a)



(b)

Figure 40: The simple model at $t = 32\,s$. (a) Trajectory of the model through phase space with the v nullcline shown in blue and the Ca^{2+} nullcline shown in green. Circles are plotted every 0.1 seconds. (b) Cartoon of the system showing Ca^{2+} and K^{+} currents as arrows with line width proportional to log of the current.

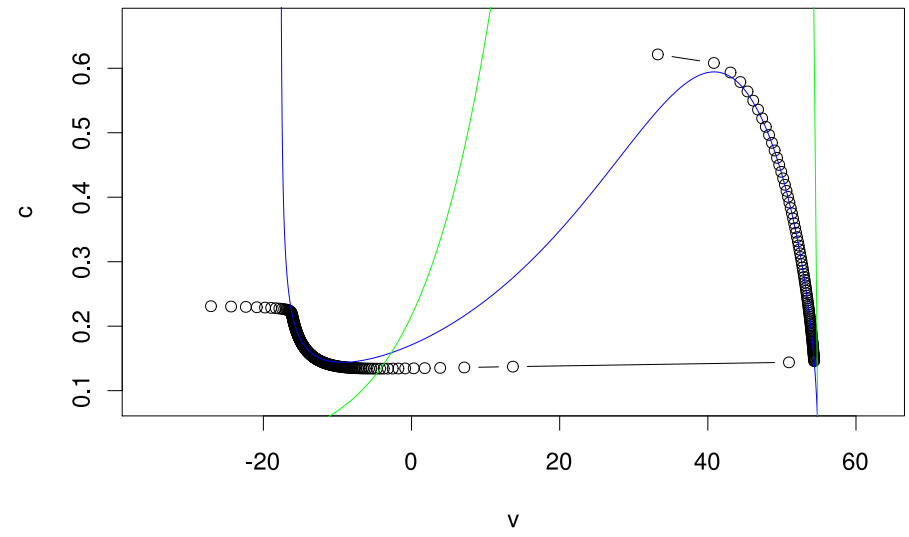
voltage are in balance and Ca^{2+} will not flow — even through a perfectly conducting channel. This behaviour of the voltage gated channel at higher voltages brings an end to the region of positive feedback ($t = 31\,s \rightarrow t \approx 32\,s$ shown in Figure 40a). The rapid rise in v comes to an end when the trajectory hits the v nullcline. The v nullcline marks the point where the Ca^{2+} activated K^+ channel is conducting enough to decrease v. In this area of phase space reducing v is important to ensure that Ca^{2+} flow through the voltage gated channel is still significant.

The region of phase space shown at $t = 32\,s$ is still at the start of a Ca^{2+} spike and is critical for whether the system oscillates or not. At some parameter values, the trajectory will hit the c nullcline before the v nullcline resulting in a drop in c to a stable fixed point. A physical interpretation of hitting the c nullcline before the v nullcline is that the Ca^{2+} pump is pumping more Ca^{2+} out of the nucleus than is flowing through the voltage gated channel which prevents a significant activation of the K^+ channel, does not decrease v and kills the resulting Ca^{2+} spike.

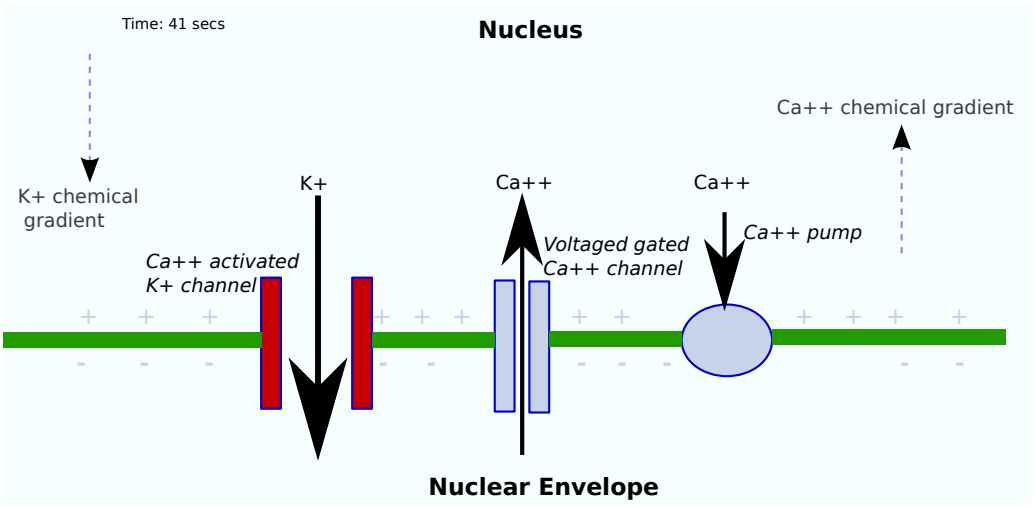
When $t = 32\,s \rightarrow t \approx 41\,s$ the trajectory of the system through phase space follows the v nullcline with increasing c and a slowly decaying v (Figure 41a). This region covers the steep rise in nuclear Ca^{2+} observed during the upward part of a spike. The increasing c improves the conductivity of the K^+ channel so that the voltage effects of i_2 slightly dominate the change in v due to i_1. At $t \approx 41\,s$ the trajectory leaves the v nullcline and v rapidly decreases which reduces the conductivity of the voltage gated Ca^{2+} channel (Figure 41b).

Closing of the voltage gated Ca^{2+} channel causes the trajectory to cross the c nullcline (Figure 42a) and the Ca^{2+} pump becomes the dominant effect. Crossing the c nullcline marks the peak of a Ca^{2+} spike. Even though the K^+ channel is highly conducting (Figure 42b) there is little movement of K^+ due to the electric field across the membrane. The trajectory then hits the v nullcline again as $i_2 \approx i_1 \approx 0$.

From $t = 41.5\,s \rightarrow t = 80\,s$ the trajectory follows the v nullcline with a slightly increasing v (Figure 43a). This region of phase space is the slow decay of a Ca^{2+} spike with the Ca^{2+} pump becoming the dominant effect and a gradually decreasing c. At $t > 80$ the trajectory follows the previous orbit around phase space and the periodic Ca^{2+} spikes continue.

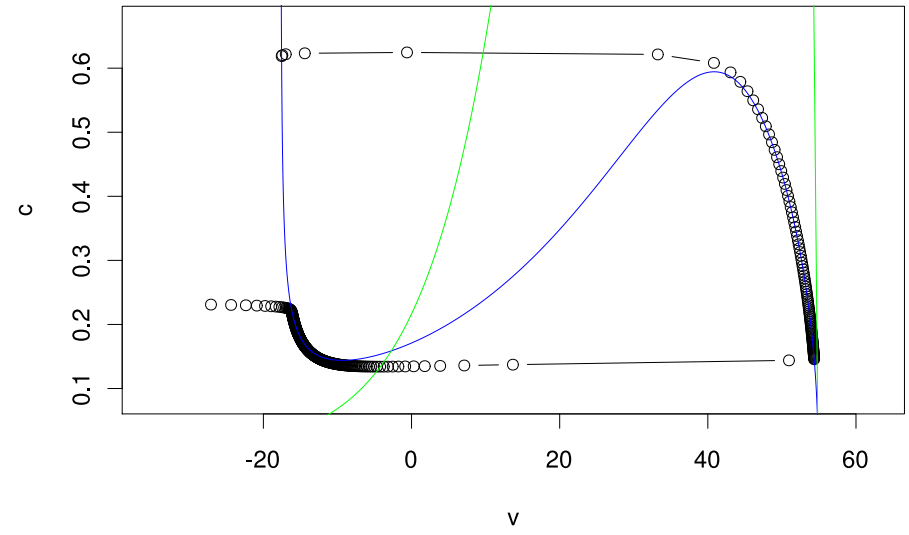


(a)

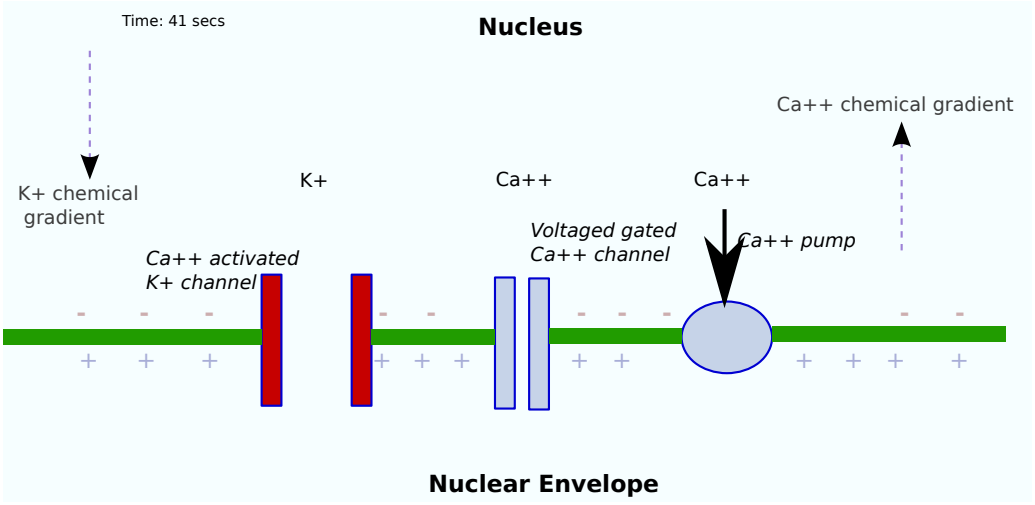


(b)

Figure 41: The simple model at t = 41 s . (a) Trajectory of the model through phase space with the v nullcline shown in blue and the Ca^{2+} nullcline shown in green. Circles are plotted every 0.1 seconds. (b) Cartoon of the system showing Ca^{2+} and K^{+} currents as arrows with line width proportional to log of the current.

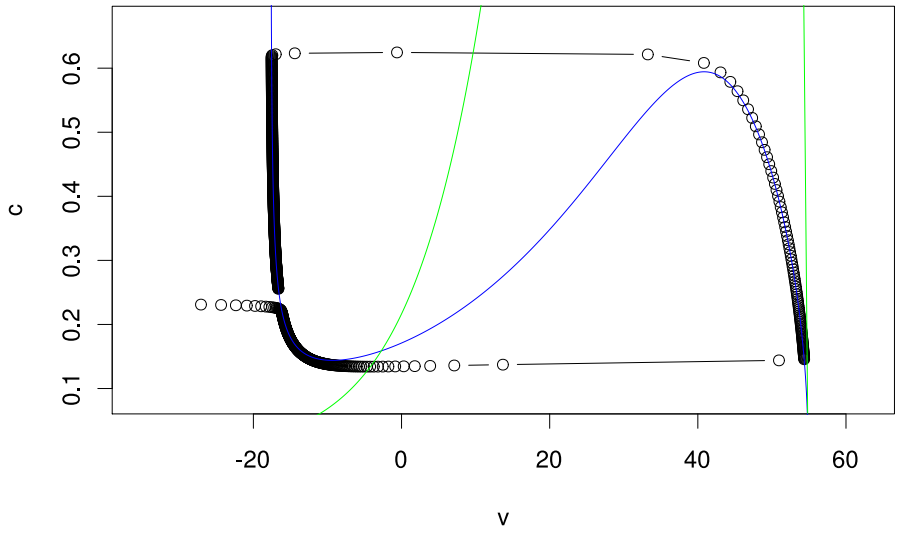


(a)

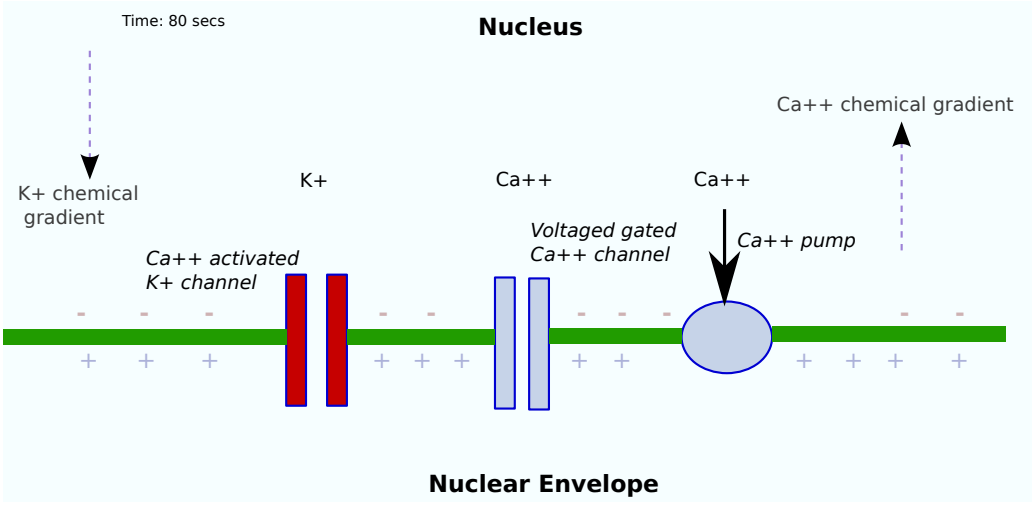


(b)

Figure 42: The simple model at $t = 41.5\,s$. (a) Trajectory of the model through phase space with the v nullcline shown in blue and the Ca^{2+} nullcline shown in green. Circles are plotted every 0.1 seconds. (b) Cartoon of the system showing Ca^{2+} and K^{+} currents as arrows with line width proportional to log of the current.



(a)



(b)

Figure 43: The simple model at t = 80 s . (a) Trajectory of the model through phase space with the v nullcline shown in blue and the Ca^{2+} nullcline shown in green. Circles are plotted every 0.1 seconds. (b) Cartoon of the system showing Ca^{2+} and K^{+} currents as arrows with line width proportional to log of the current.

4.4 MODELLING CA^{2+} PROBES AND BUFFERS

Now that a periodic spiking model exists, modifications can be made to investigate experimental and physiological effects. The concentration of Ca^{2+} buffers in the nucleus and the concentration of the dye or probe used to measure Ca^{2+} are two potential influences on Ca^{2+} spiking. Each of these sources of Ca^{2+} buffering could be modelled using similar equations to the ones used by Marhl et al. [77]. However, because the rate constants, k_- and k_+, are not known for the Ca^{2+} probe or nuclear Ca^{2+} binding proteins, we use dissociation constants, $k_d = \frac{k_-}{k_+}$, and a fast buffer approximation.

Consider k_b the dissociation constant for a Ca^{2+} buffering protein with a total concentration b_{tot}, of which a concentration, b, is bound to Ca^{2+},

$$k_b = \frac{(b_{tot} - b)c}{b} \tag{4.6}$$

$$b = \frac{b_{tot}c}{k_b + c}\,, \tag{4.7}$$

where c is the concentration of free Ca^{2+}. From equation 4.7, the rate of change of occupied buffer with respect to free calcium can be obtained,

$$\frac{db}{dc} = \frac{k_b b_{tot}}{(k_b + c)^2}\,. \tag{4.8}$$

Using the same process, p, the concentration of Ca^{2+} probe bound to Ca^{2+} can be written in terms of its total concentration p_{tot} and its dissociation constant, k_p,

$$\frac{dp}{dc} = \frac{k_p p_{tot}}{(k_p + c)^2}\,. \tag{4.9}$$

The effects of Ca^{2+} probe and nuclear Ca^{2+} buffering proteins are then incorporated into Equation 4.2,

$$\frac{dc}{dt}\left(1 + \frac{dp}{dc} + \frac{db}{dc}\right) = \alpha i_1 - \mu c\,. \tag{4.10}$$

The nature of the Ca^{2+} buffers in the nucleus is not known. However, as an approximation for k_b the dissociation constant of the ubiquitous Ca^{2+} binding protein calmodulin is used [63]. The addition of Ca^{2+} buffering requires the unknown parameters to be re-estimated as shown in Table 6. Most values are similar with the notable exception of the resting potential for K^+, E_k, which has dropped from 17.7 mV to 8.8 mV.

Due to an inflexibility in the parameter estimation implementation, the buffered model was fitted using c rather than p. Comparing

Table 6: Parameter values used in the buffered model with their sources. The source 'Fit' indicates the parameter value was obtained by fitting the model to a Ca^{2+} spike

	Description	Value	Units	Source
k_b	Dissociation constant of Ca^{2+} buffering proteins in the nucleus	1.0	μM	[63]
k_p	Dissociation constant of Ca^{2+} probe when situated in the nucleus	0.32	μM	[123]
b_{tot}	Concentration of Ca^{2+} buffering proteins in the nucleus	943	μM	Fit
p_{tot}	Concentration of Ca^{2+} probe	544	μM	Fit
G_c	Total max conductance of voltage gated Ca^{2+} channels	2872	pS	Fit
V_{ml}	Half maximal activation of voltage gated Ca^{2+} channel	50.0	mV	Fit
K_{ml}	Constant in scaling function for voltage gated Ca^{2+} channel	13.2	mV	Fit
G_k	Total max conductance of Ca^{2+} activated K^+ channels	278	pS	Fit
K	Constant in Hill function for Ca^{2+} activated K^+ channel	0.879	μM	Fit
μ	Pump rate into the nuclear envelope	22.2	s^{-1}	Fit
E_k	Resting potential of K^+	8.8	mV	Fit

normalised time series for c and p (Figure 44) suggests that fitting to c is not strictly correct and that parameter values will be different if it is recognised that the time series record the response of a Ca^{2+} probe rather than actual Ca^{2+} concentration.

4.4.1 *Sensitivity Analysis*

In order to understand the importance of Ca^{2+} buffers and measure the significance of parameter changes on the model, a local sensitivity analysis was conducted. This technique perturbs the values of parameters and measures the effects of the perturbation on the integrated

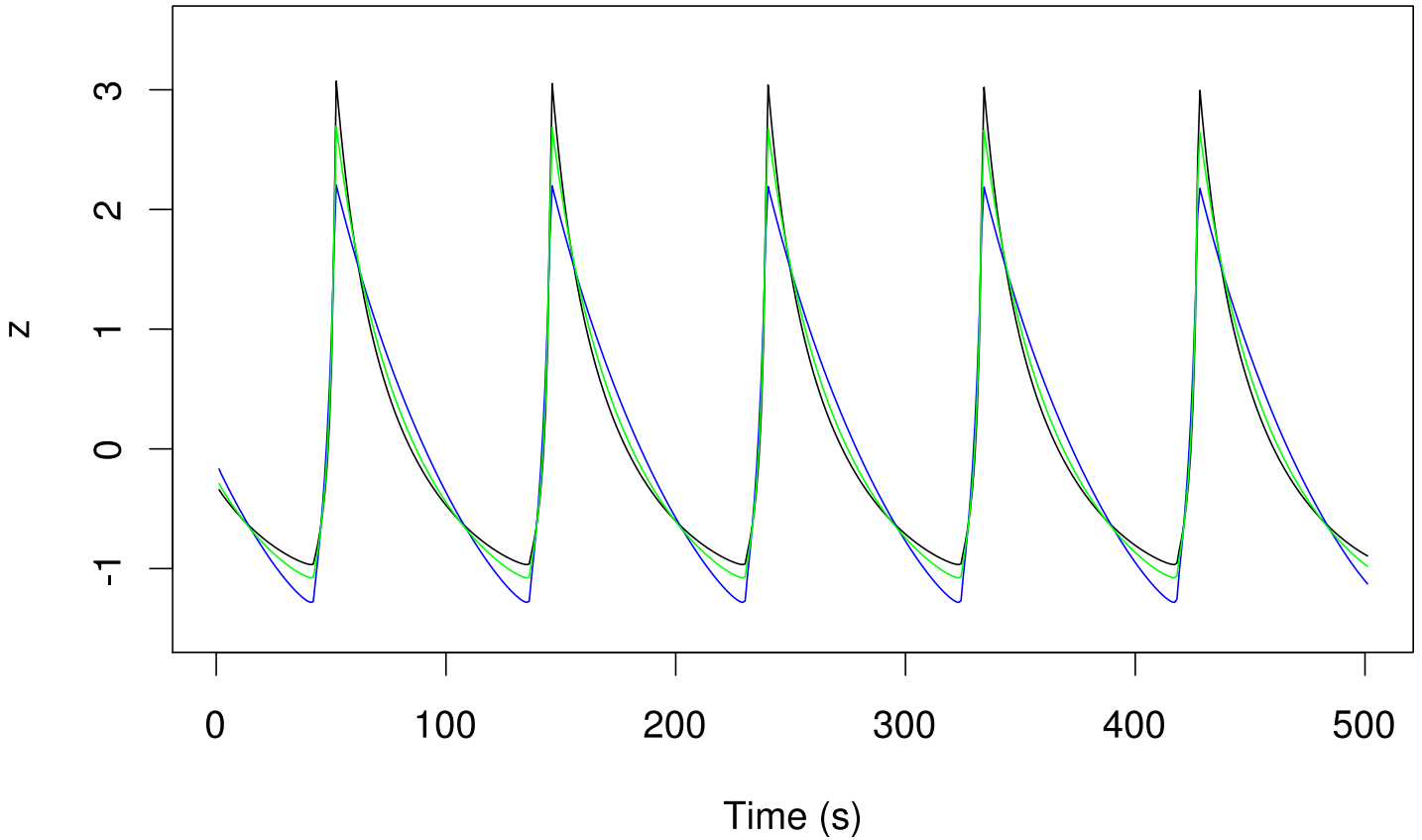


Figure 44: Ca^{2+} spikes and Ca^{2+} binding as generated by the model given in Equation 4.10. The y axis indicates the normalised time series, $z = \frac{x-\mu}{\sigma}$, where x is a time series measurement, μ is the mean of a time series and σ is the standard deviation of a time series. Free Ca^{2+} is shown in black, Ca^{2+} bound to the Oregon Green indicator in blue and Ca^{2+} bound to buffer proteins in the nucleus is denoted with green.

Table 7: Normalised output values, to one decimal place, for the effects on baseline, amplitude and period of Ca^{2+} spikes when perturbing parameter values. The output values were calculated using equation 4.11.

	baseline	amplitude	period
V_{ml}	-4.4	-3.3	2.6
G_k	-0.5	-0.7	0.3
K_{kc}	0.9	1.0	-0.6
G_{ca}	0.6	0.7	-0.3
μ	-0.1	0.0	-1.0
E_k	-0.8	0.1	0.5
K_{ml}	4.6	-0.4	-3.4
p_{tot}	0.0	-0.0	0.2
b_{tot}	0.0	-0.1	0.7

output. The sensitivity analysis was carried out using a normalised centred difference approximation [131],

$$S_{ij} = \frac{\frac{O_i(p_j + \Delta p_j) - O_i(p_j - \Delta p_j)}{O_i}}{\frac{2\Delta p_j}{p_j}}, \tag{4.11}$$

where p_j is the j-th parameter which is perturbed, $\Delta p_j = 0.001 \times p_j$, to produce model output O which can have multiple components indexed as O_i.

Conventionally, O_i is the squared difference in the i-th sample of the integrated time series. However, because the Ca^{2+} spiking model is oscillating, comparing time series data points may overemphasise the significance of changes to frequency or phase. This is a shortcoming that is similar to using least squares for parameter estimation as discussed in Section 3.3.1. As alternative model outputs, we measured the basal Ca^{2+} concentration, the maximal Ca^{2+} concentration and the period between Ca^{2+} spikes.

The model is particularly sensitive to changes in the voltage gated Ca^{2+} channel with the parameters V_{ml}, the half maximal activation voltage, and K_{ml}, the scaling function constant being the most sensitive parameters in the model (Table 7). The model is not sensitive to the concentration of Ca^{2+} probe, p_{tot}, suggesting that small changes in dye concentrations will not significantly effect Ca^{2+} spiking. However, because the order of the magnitude of p_{tot} differences is not known

when performing experiments, the concentration of Ca^{2+} dye could still be a cause of experimental variability.

4.5 CALCIUM INDUCED CALCIUM RELEASE

The model described in Section 4.3 releases Ca^{2+} in response to voltage changes across the membrane of the nuclear envelope. Even though the model is not spatial, it is possible to determine some of qualities that a spatial form of the model would have. Voltage changes move rapidly across membranes and this suggests that all voltage gated channels would release Ca^{2+} at roughly the same time.

Using a nuclear localised Ca^{2+} indicator and confocal imagery, Sieberer et al. [117] have shown that Ca^{2+} is released in the form of puffs into the nucleoplasm. The puffs do not occur simultaneously and different locations release Ca^{2+} as far as 5 seconds apart. Section 4.6 discusses a way in which non-uniform release could be possible with voltage activated Ca^{2+} channels. However, a less contentious way of modelling this behaviour is with an activator that diffuses to fire Ca^{2+} channels at different time points.

A well known activator in animal systems is Ca^{2+} itself which operates through a positive feedback process known as Calcium Induced Calcium Release (CICR). Channels release Ca^{2+} which diffuses to nearby channels and activates them causing further Ca^{2+} release. CICR can be incorporated into Equation 4.4 by augmenting the voltage gated Ca^{2+} channel with Ca^{2+} activation in the form of a Hill function,

$$i_1 = G_c f(v) \frac{c^2}{c^2 + K_{cicr}^2} (E_{ca} - v) \tag{4.12}$$

Starting with the parameter values given in Table 5, some modifications were made. First, the effects of the CICR term were removed by setting K_{cicr} to 0 μM. Then spontaneous spiking was disabled by raising K_{ml} to 17 mV. The effects of CICR were enabled by setting K_{cicr} to 0.1 μM. With CICR enabled in this way, spontaneous spiking is restored (Figure 45).

4.6 DISCUSSION

This chapter describes deterministic models of 2 dimensions that are built on assumptions and attempt to capture the qualities of a more complex system. It is worth examining the assumptions, to see how rational they are, and to suggest alternatives.

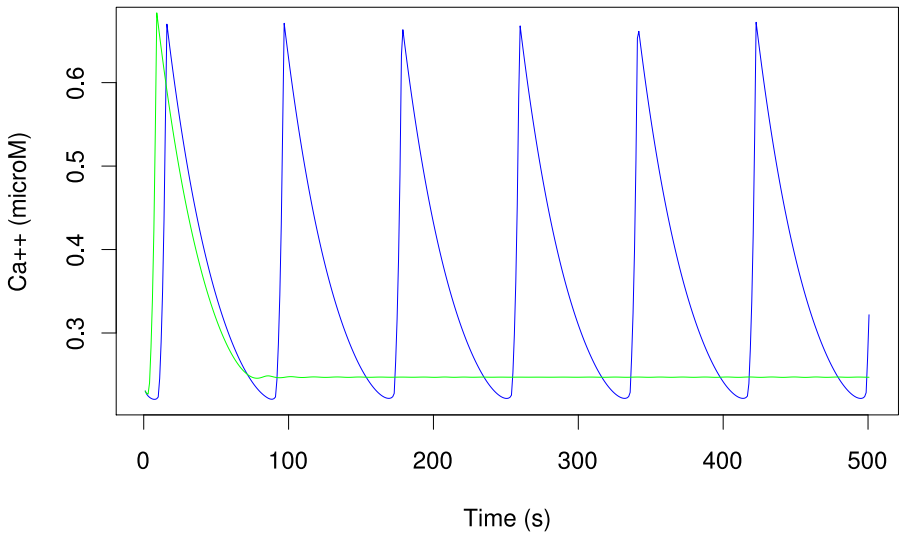


Figure 45: Ca^{2+} spikes generated by the model given in Equation 4.12 with CICR enabled (blue). With CICR removed spiking does not persist (green).

Taking the volume of the nucleus to be 160 fL (Table 5), leads to an estimated 9635 calcium ions in the nucleus at a basal Ca^{2+} concentration of 100 nMol. Using a rough estimate that uncertainties due to low copy number are of the order of $\frac{1}{\sqrt{N}}$, where N is the number of particles, gives a 1% stochastic fluctuation. This indicates that deterministic dynamics will still be significant enough that modelling with differential equations is a valid approach to contribute to the understanding of nuclear Ca^{2+} spiking. However, it would be instructive to do a stochastic simulation to see how the model behaves around the critical fixed point that effects the initiation of a Ca^{2+} spike (Figure 40).

In the current models, the Ca^{2+} pump does not contribute directly to the inner nuclear membrane potential. An electrogenic Ca^{2+} pump would contribute directly to changes in the membrane potential by producing a net charge transport. This alteration to the models would reduce the required conductivity of the K^+ channels that currently provide a balance. The K^+ channels are already modelled with a small conductance, 302 pS for the whole nucleus. The Ca^{2+} pump is modelled as an exponential term that increases Ca^{2+} transport without limits as the free Ca^{2+} concentration increases. Converting

this pump equation to a Hill function to model co-operative binding and a saturation of pumping rate could attenuate the effects of an electrogenic transport on membrane potential.

The voltage gated Ca^{2+} channel is a hypothesised channel whose very existence can be doubted. The dynamics of the channel described by Equation 4.3 are therefore arbitrary. The channel is essential for the described models though for 2 reasons. Genetically, DMI1 is required for Ca^{2+} spiking and a reasonable assumption is that the transport of alternative cations indicate a need to balance membrane potential. A mechanism must release Ca^{2+} from the nuclear envelope and also cause the Ca^{2+} channel to close at the peak of a spike. A voltage mechanism is a parsimonious premise to make since alternatives, such channels closing due to Ca^{2+} inhibition, add additional assumptions and parameters to the model.

An intracellular membrane potential, a voltage across the ER, has been modelled as part of Ca^{2+} oscillations in animal systems [77]. In the animal model, electroneutrality was considered along with anion, cation and free buffer concentrations. The animal model was more encompassing than the ones described in this analysis. It would be instructive to see if a similar approach would work with symbiotic Ca^{2+} spiking.

Another obvious simplification of the models here is the lack of an active transport for K^+. Although the K^+ current required to balance membrane potential is small enough not to affect nuclear K^+ concentration, the active transport of K^+ back into the nucleus may affect membrane potential. However, because the transporters for K^+ are not known, their inclusion would only complicate the model.

The models described here predict a concentration gradient of K^+ between the nuclear envelope and the nucleus. Compared to Ca^{2+}, relatively little is known about the role of K^+ within the nucleus, or even the cytosol, of plant cells. Cells keep a high internal K^+ concentration and intracellular indicators for this ion would be of little use. However, in animal systems there are suggestion that there is a K^+ gradient across the membrane of the nuclear envelope because a voltage and Ca^{2+} activated K^+ channel has been isolated from pancreatic acinar cells [122]. Knowing that K^+ plays a major role in symbiotic Ca^{2+} oscillations and not being able to measure the behaviour of K^+ during the oscillations will probably be a continuing difficulty in modelling the symbiotic Ca^{2+} spiking system.

Voltage induced Ca^{2+} release may suggest simultaneous Ca^{2+} release from all channels on a membrane. Sluggish release of Ca^{2+}

has been observed, however, in images of nuclear Ca^{2+} spiking [117]. A likely explanation is another release mechanism such as CICR is involved. An alternative interpretation is that voltage gated channels on the inner nuclear membrane have a non-uniform voltage response with some channels activating at different voltages to others. The voltage gated model analysed in Section 4.4.1 is particularly sensitive to V_{ml} and K_{ml} — the parameters that describe how Ca^{2+} transport through the voltage gated channel is affected by membrane voltage. A more radical explanation is that no Ca^{2+} is released in the nucleus at all and that the observed Ca^{2+} puffs in [117] have arrived from the cytosol after travelling through nuclear pores which may modulate the transport of Ca^{2+} [79]

4.7 SOFTWARE USED

Exploratory analysis on candidate models was performed using XPPAUT [26]. The integrated data from XPPAUT was used to animate SVG drawings using a Python program. The output of the Python program was converted into an animated film and still images (Figures 38 to 43). Parameter estimation was done using the particle swarm optimiser described in Chapter 3, Algorithm 1. Phase space plots were produced by a program written in R that used the odesolve library to integrate the models. Sensitivity analysis was also performed by a program written in R using a description of the local sensitivity algorithm given in [131].

5

CONCLUDING REMARKS

5.1 CONCLUDING REMARKS

At the time of writing no mathematical models of perinuclear Ca^{2+} spiking in plants have been published. However, Brière et al. [15] have developed a model of Ca^{2+} release after mechanical stimulation of plant nuclei. The absence of models is probably due, in part, to the lack of knowledge on individual components. Not knowing the critical components in an oscillating system with a chaotic signature is a modelling challenge but it is also an opportunity to contribute to the work being done in biology labs to understand symbiotic Ca^{2+} spiking.

This work has used three different computational approaches that attempted to extract information about the symbiotic Ca^{2+} spiking that occurs in *Medicago truncatula* during a symbiosis with nitrogen fixing bacteria. The first approach was a nonlinear time series analysis, that although not conclusive, suggests that irregularities in the Ca^{2+} oscillations are not predominately stochastic in nature and that the states of the system could orbit a chaotic attractor. Although this could be seen as a labelling exercise it is still part of a wider attempt to understand the system being studied. The time series analysis has implications for models of the Ca^{2+} spiking. The modelling described in this work has been deterministic. However, if the alternative approach of developing a stochastic model is taken, the resulting model could be analysed to see if it produced time series with the same nonlinear characteristics as the experimental data.

The second approach was to build models computationally by fitting to time series. It was shown that it is possible to search for viable models using Inductive Process Modelling but that a selection of available parameter estimation techniques were unable to identify parameters accurately enough for model discrimination. Although not usable on a chaotic system, the framework discussed could have applications in other areas of systems biology as shown by a successful demonstration with a periodically oscillating Ca^{2+} model. This systems identification approach was the least successful at improving our understanding of symbiotic Ca^{2+} spikes. However, it also has the most potential

since it allows hypotheses to be made about components and then have those hypotheses scored with regards to how well they explain experimental data and how parsimonious they make the resulting models. Nonlinear time series analysis doesn't allow determination of individual components while manual modelling does not usually incorporate a step to score one set of assumptions in comparison to another.

When modelling the symbiotic Ca^{2+} spiking system using conventional techniques only a relatively small number of models and assumptions can be examined. Without an exhaustive search of model space this could been seen as an insurmountable problem. One hopeful comparison can be to the One Pool model of Ca^{2+} oscillations in animal systems [21]. At the time the One Pool model was developed, the Ca^{2+} store and the behaviour of the IP_3 receptor had not been fully characterised. This led to the development of a model that used parsimonious assumptions to fill in missing details and proposed a mechanism for Ca^{2+} spiking that is noticeably close to the one recognised in animal systems today. This in no way suggests that the predictions made by the models in Chapter 4 will be as accurate of those produced by the One Pool model. However, the history of the One Pool model demonstrates that models that rely on hypothesised Ca^{2+} channels may have potential utility.

Possible future work could follow the analysis done for this thesis. One course of action is to investigate the use of Inductive Process Modelling (IPM) on other biological systems. Currently, a periodic model, based on the simplest of assumptions, fits well to a single Nod Factor induced Ca^{2+} spike. It is unlikely that applying IPM to the experimental data would suggest an alternative model. However one possible application for IPM is to look at the Ca^{2+} oscillations that occur in animal systems which have many competing models which have not been fit to experimental data.

Since only a periodic model exists for Nod Factor spiking in *M. truncatula*, it would be interesting to see if a viable chaotic model can be found. There are two main approaches that could be taken. The first is to add variables to the current models to investigate if a mechanism can be identified that induces chaos in a system of ODEs. The alternative is to model the system spatially to see if the behaviour of the Ca^{2+} oscillations could be the result of spatial chaos.

If a chaotic model can be established for the system, it would be natural to use this model to suggest a parameter that can be analysed for bifurcations. This could be performed mathematically

and possibly experimentally. Finding a parameter of the biological system that can be used to take the Ca^{2+} spikes from periodic to chaotic behaviour would convincingly demonstrate that the system is chaotic. An experiment of this type has shown that population dynamics can be chaotic [7].

Part II

APPENDICES

1

NOD FACTOR TRACES

1.1 OVERVIEW

This appendix contains plots of the experimental data analysed in Chapter 2. The time series Nod1 is plotted in the main text in Chapter 2 as Figure 7.

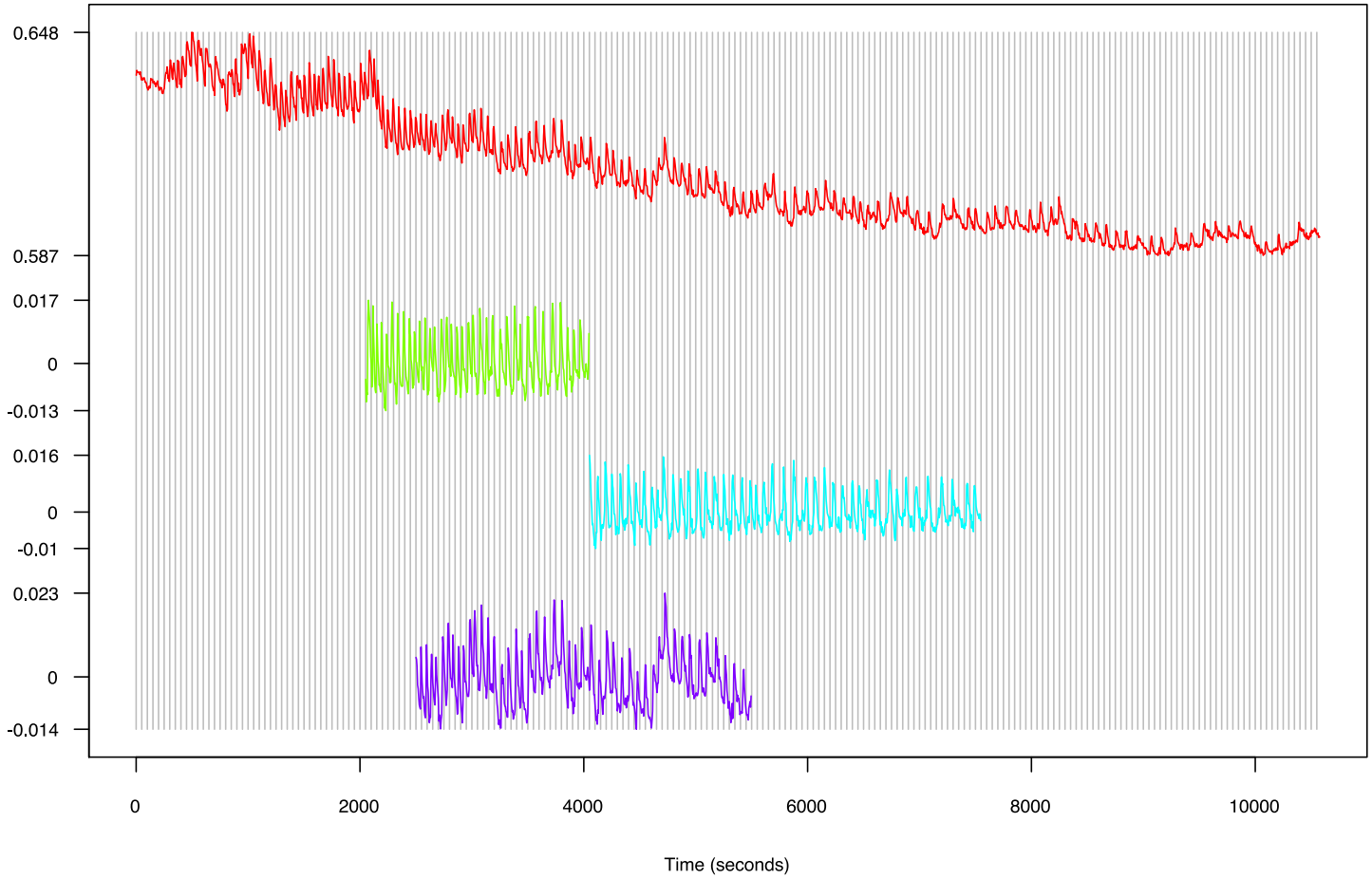


Figure 46: An original experimental time series (red), after detrending with a moving average to produce Nod2 (green) and Nod3 (cyan). After detrending with EMD only trace Nod4 (purple) could be extracted under the contraint that the time series has to be stationary. The Y axis is a fluorescence ratio between Ca^{2+} sensitive and Ca^{2+} insensitive dyes. The X axis is time in seconds.

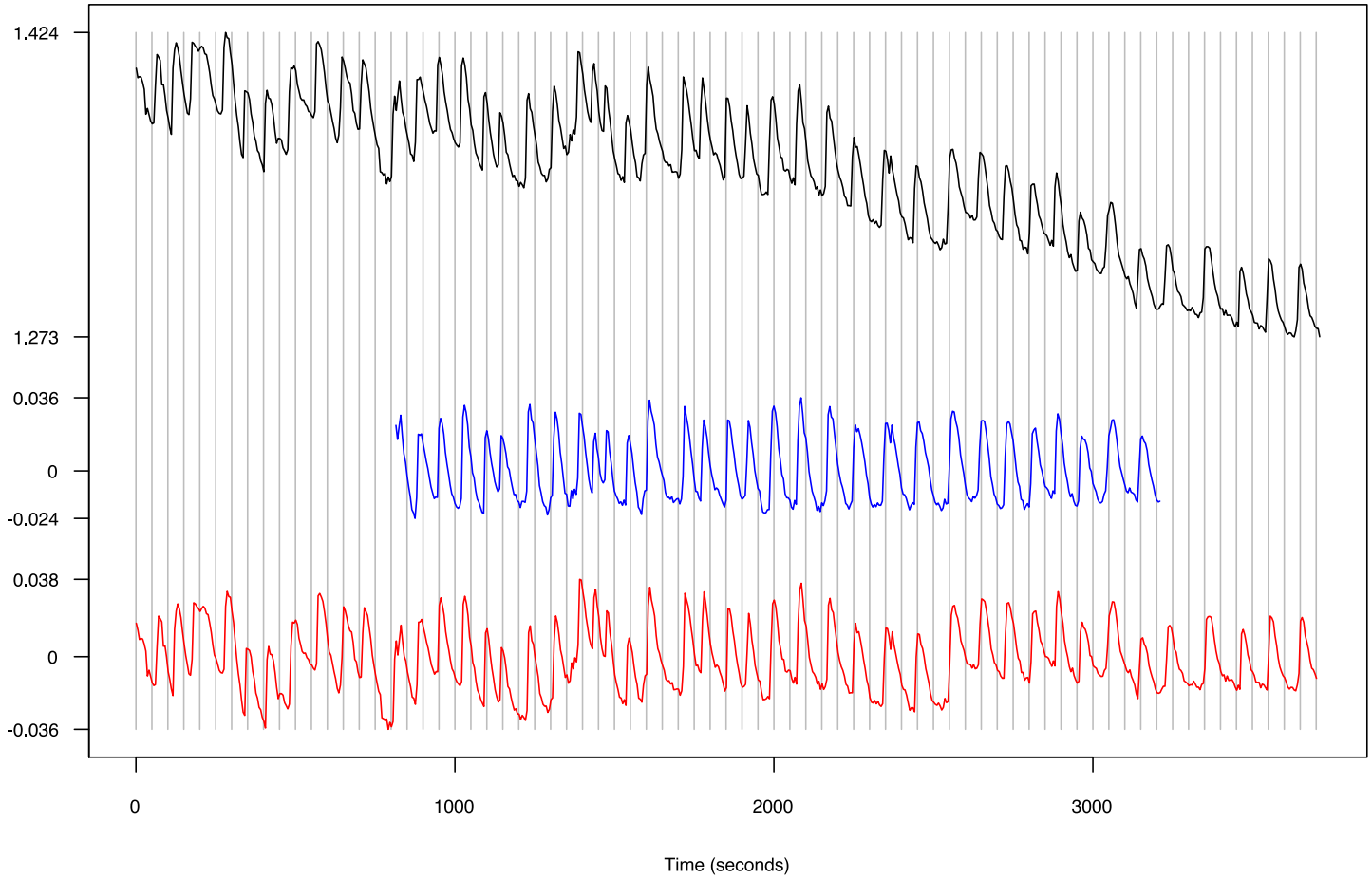


Figure 47: The Nod5 time series (black), after detrending with a moving average (blue) and after detrending with EMD (red). The Y axis is a fluorescence ratio between Ca^{2+} sensitive and Ca^{2+} insensitive dyes. The X axis is time in seconds.

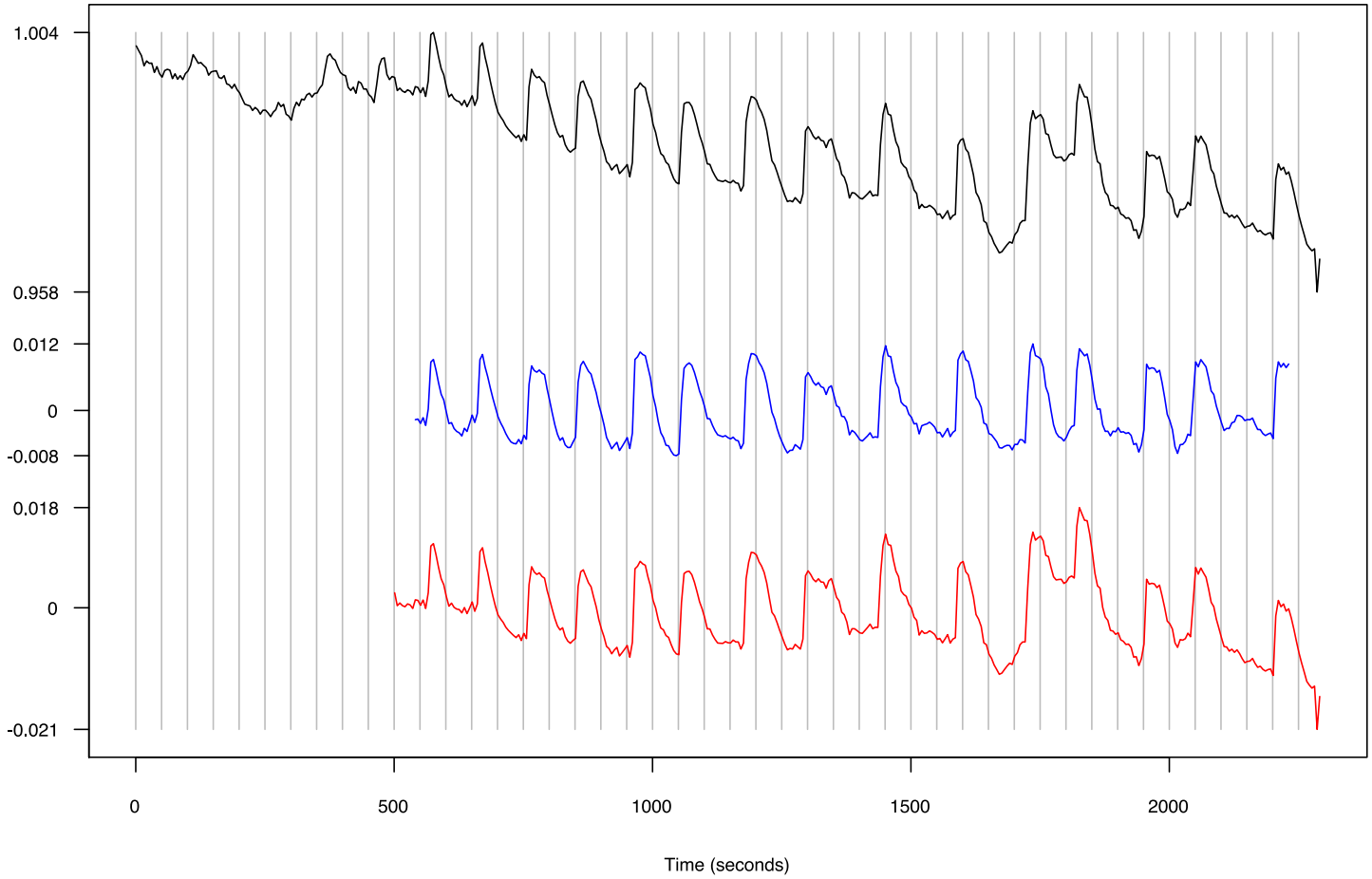


Figure 48: The Nod6 time series (black), after detrending with a moving average (blue) and after detrending with EMD (red). The Y axis is a fluorescence ratio between Ca^{2+} sensitive and Ca^{2+} insensitive dyes. The X axis is time in seconds.

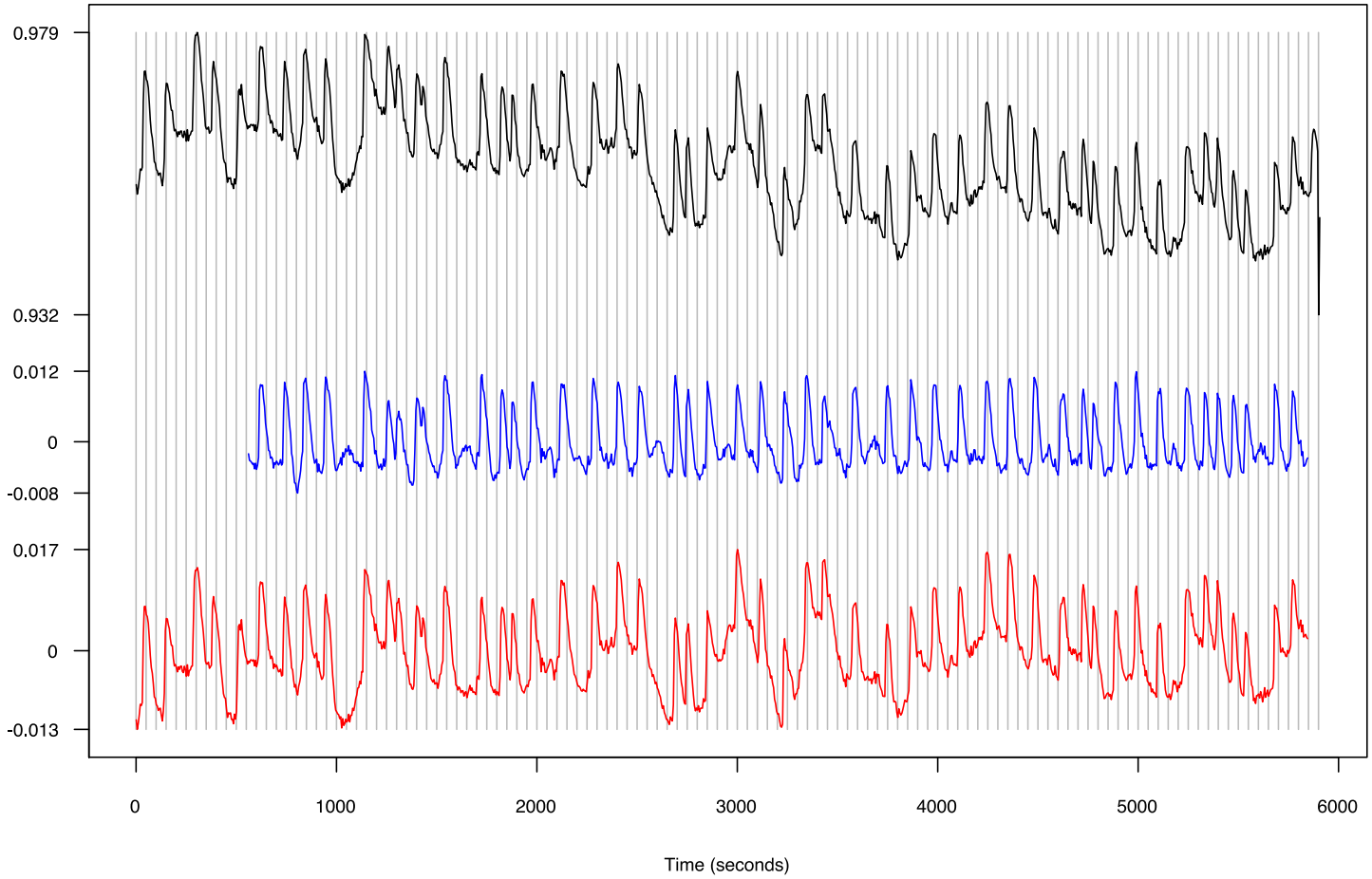


Figure 49: The Nod7 time series (black), after detrending with a moving average (blue) and after detrending with EMD (red). The Y axis is a fluorescence ratio between Ca^{2+} sensitive and Ca^{2+} insensitive dyes. The X axis is time in seconds.

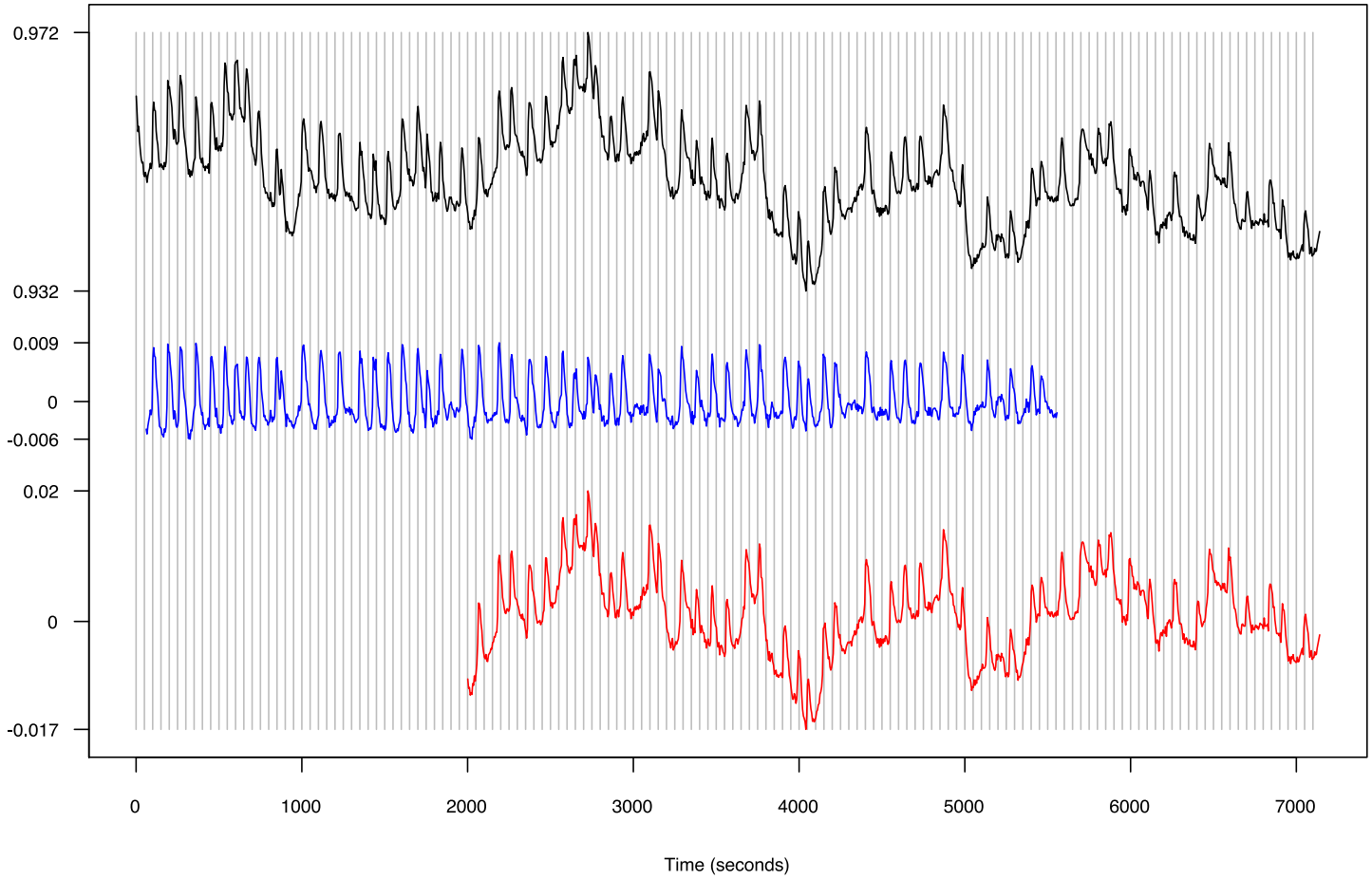


Figure 50: The Nod8 time series (black), after detrending with a moving average (blue) and after detrending with EMD (red). The Y axis is a fluorescence ratio between Ca^{2+} sensitive and Ca^{2+} insensitive dyes. The X axis is time in seconds.

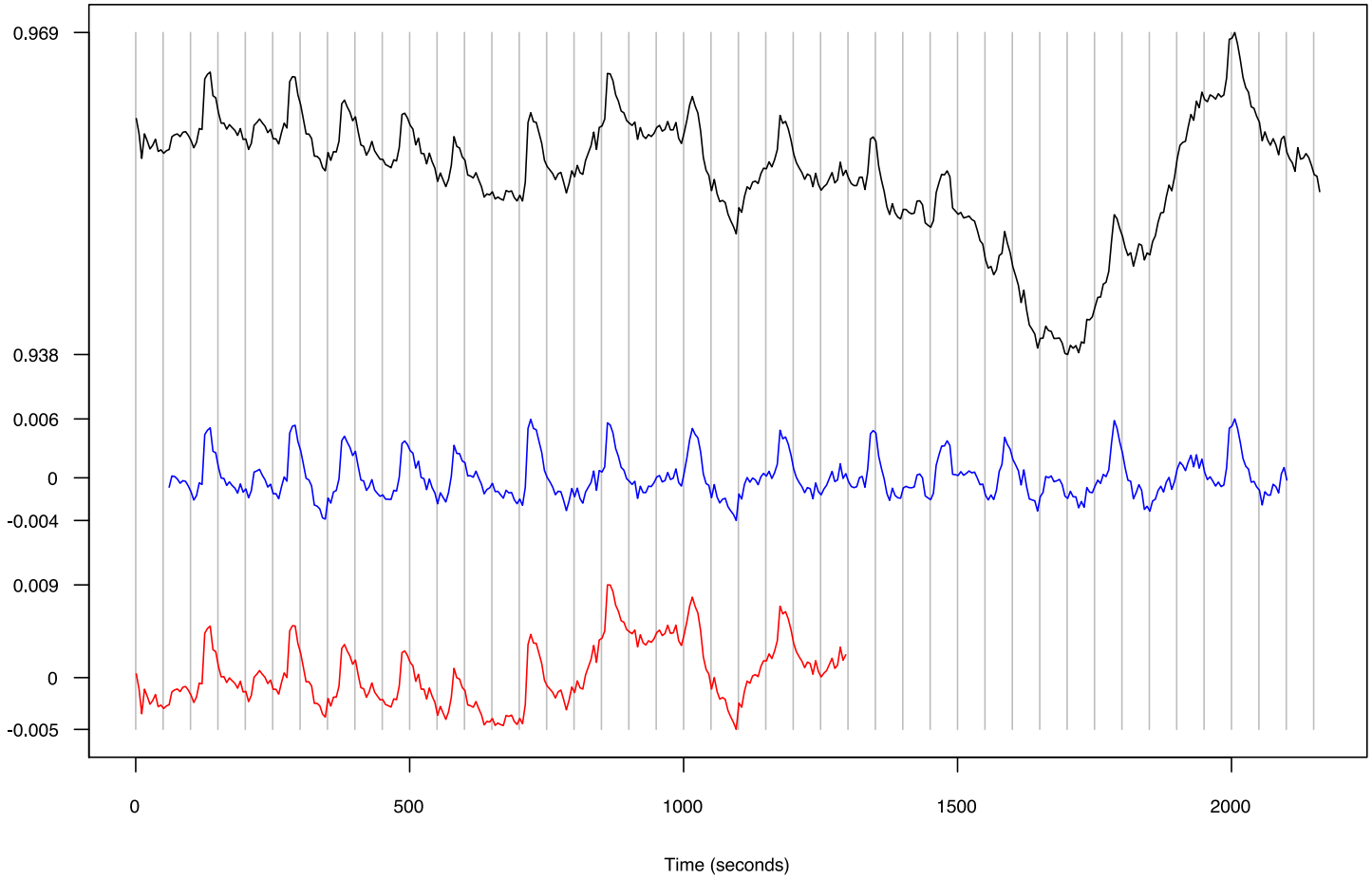


Figure 51: The Nod9 time series (black), after detrending with a moving average (blue) and after detrending with EMD (red). The Y axis is a fluorescence ratio between Ca^{2+} sensitive and Ca^{2+} insensitive dyes. The X axis is time in seconds.

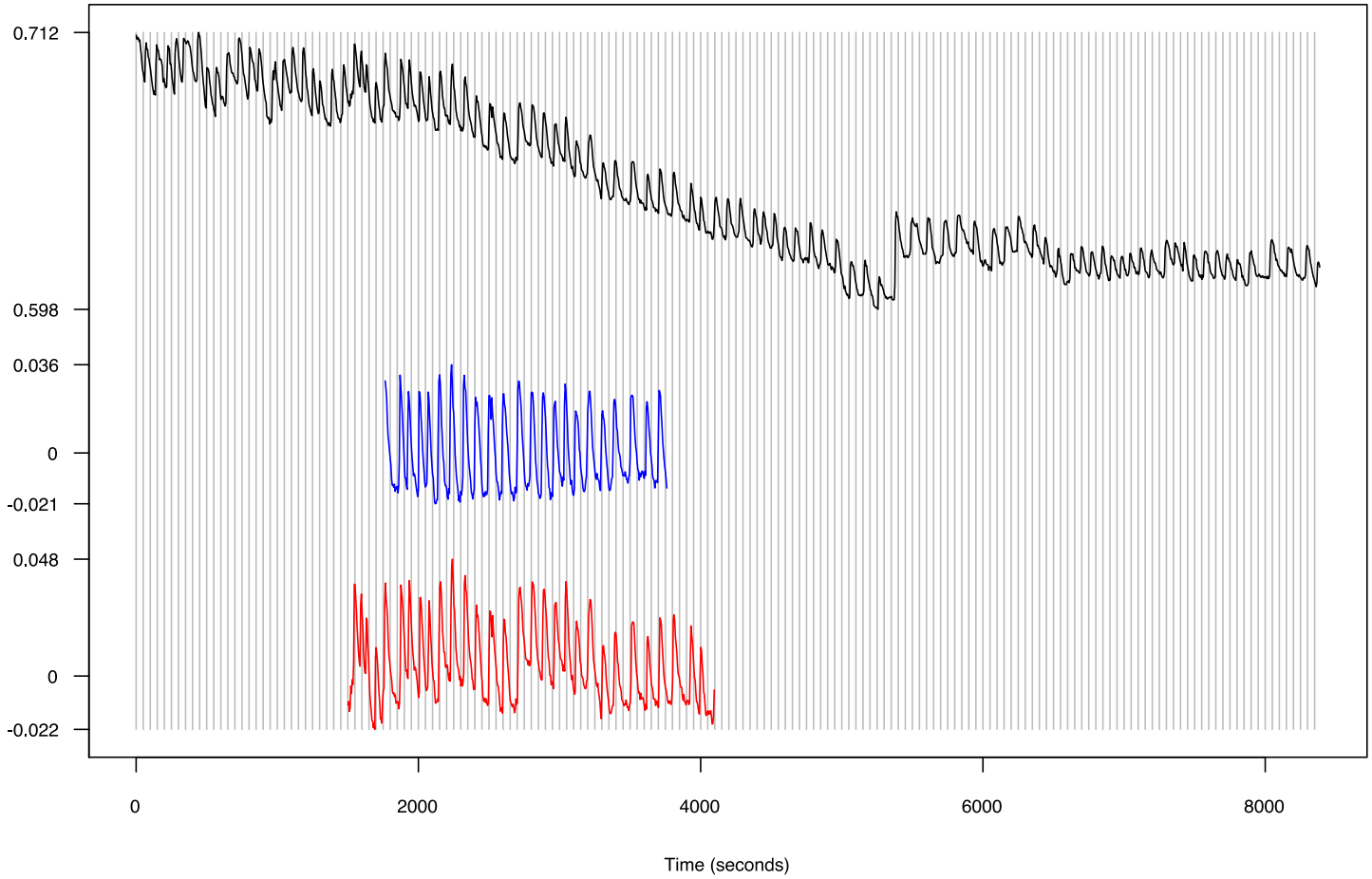


Figure 52: The Nod10 time series (black), after detrending with a moving average (blue) and after detrending with EMD (red). The Y axis is a fluorescence ratio between Ca^{2+} sensitive and Ca^{2+} insensitive dyes. The X axis is time in seconds.

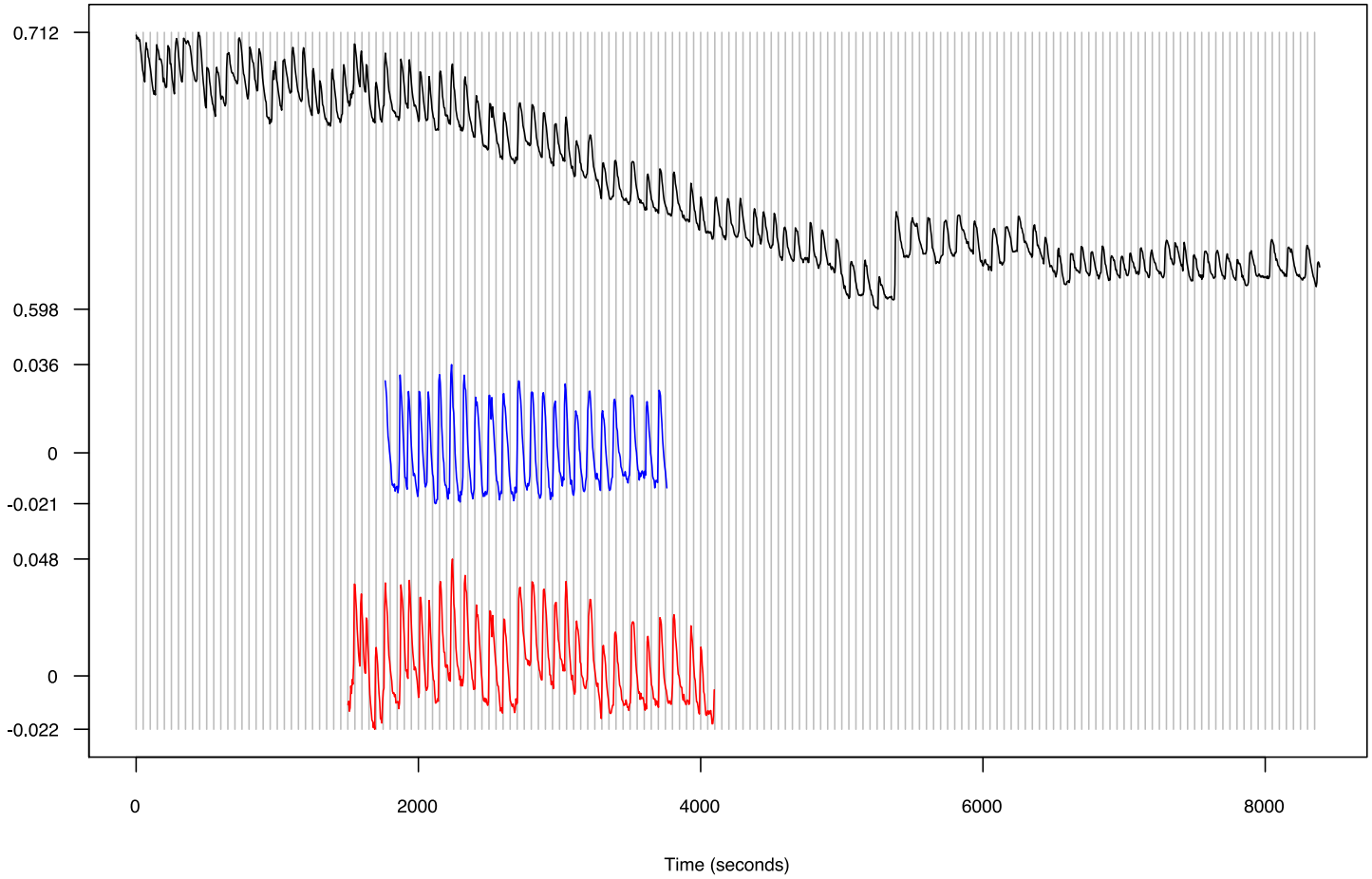


Figure 53: An original experimental time series (black), after detrending with a moving average (Nod11 in blue and Nod12 in cyan) and after detrending using EMD (Nod11 in red and Nod12 in purple). The Y axis is a fluorescence ratio between Ca^{2+} sensitive and Ca^{2+} insensitive dyes. The X axis is time in seconds.

2

INDUCTIVE PROCESS MODELLING

Below is a listing in Python that captures the Equations 3.32 to 3.42 in Chapter 3. It uses an inductive modelling library that defines the classes `DAE`, `CFunction`, `Variable` and `Parameter`. These defined classes have overloaded operators `+ - / * |` that allow equation and modelling objects to be built using an equation-like syntax.

The `|` operator indicates equation alternatives and is equivalent to the $\bigvee$ operator in Section 3.5.2.

```
from inductive_modelling import *

class Hill(CFunction) :
    name = "hill"

class Power(CFunction) :
    name = "pow"

class Schuster(DAE) :
    """
        Minimal models taken from table 1 of
        Modelling of simple and complex calcium oscillations
        Stefan Schuster et al
        European Journal Biochemistry 269, 1333-1355, 2002

        96 alternative models of which 24 are valid
    """

    Ca_cyt  = Variable(0, 2.0, 1.0)
    Ca_er   = Variable(0, 2.0, 1.0)
    B       = Variable(0, 1000, 500)
    R       = Variable(0, 1.0, 0.5)

    # Parameters common to all models
    rho_er  = Parameter(0.9,5,1)

    # Parameters from Dupont and Goldbeter converted to seconds
    k_f     = Parameter(0,0.05,0.017)
    k       = Parameter(0,0.5,0.17)
    v0      = Parameter(0.01,0.1,0.08)
    v1      = Parameter(0.01,0.1,0.08)
    beta    = Parameter(0,0.5,0.1)
    VM2     = Parameter(0,1.0,0.83)
    K_2     = Parameter(0.4,2,1)
    VM3     = Parameter(1,20,6.8)
    K_R     = Parameter(1,3,2)
    K_A     = Parameter(0.4,1,0.5)

    # Parameters from Li and Rinzel
```

```
    k0      = Parameter(0,0.1,0.02)
    k1      = Parameter(10,50,40)
    K_a     = Parameter(0,1,0.4)
    k_3     = Parameter(0,1,0.2)
    k__3    = Parameter(0,1,0.5)

    # Parameters from Marhl et al
    k_leak  = Parameter(1,20,10)
    k_ch    = Parameter(0,1,0.6)
    K_1     = Parameter(1,10,5)
    k_pump  = Parameter(1,100,76)
    k_plus  = Parameter(0,1,0.1)
    B_0     = Parameter(100,1000,600)
    k_      = Parameter(0,1,0.5)

    # Rate laws
    Vin     = v0 + v1 * beta | 0
    Vout    = k * Ca_cyt
    Vrel    = k_f * Ca_er + beta * VM3 * Hill(Ca_er, K_R, 2) * Hill(Ca_cyt, K_A, 4) | \
              (k0 + k1 * R * Power(Hill(Ca_cyt, K_a, 1),3)) * (Ca_er - Ca_cyt) | \
              (k_leak + k_ch * Hill(Ca_cyt, K_1, 2)) * (Ca_er - Ca_cyt)
    Vserca  = VM2 * Hill(Ca_cyt, K_2, 2) | \
              k_pump * Ca_cyt
    Vrec    = k_3 * (1 - R)
    Vdes    = k__3 * Ca_cyt * R
    Vb      = k_plus * (B_0 - B) * Ca_cyt - k_ * B

    # Differential equations
    d = {}
    d[Ca_cyt] = Vin - Vout + Vrel - Vserca - Vb | \
                Vin - Vout + Vrel - Vserca
    d[Ca_er]  = rho_er * (Vserca - Vrel)
    d[B]      = Vb | None
    d[R]      = Vrec - Vdes | None



eq_system = Schuster()
```

These definitions are processed and used to generate multiple C++ header files, each defining a system of equations. An example header file is shown below with line breaks added for clarity.

```
#ifndef AUTO_MODEL_H
#  define AUTO_MODEL_H

#include "simple_model.h"


class AutoModel : public SimpleModel
{
   public:
      enum Var_name   {  Ca_cyt ,  Ca_er ,  VAR_END };
      enum Param_name {  K_R ,  k ,  K_2 ,  v0 ,  v1 ,  beta ,  K_A ,
                         rho_er ,  VM2 ,  VM3 ,  k_f ,  PARAM_END };

      AutoModel() :
```

```
        SimpleModel("AutoModel", VAR_END, PARAM_END)
    {
        SimpleModel::add_var(Ca_cyt,"Ca_cyt",0,2.0,1.0);
        SimpleModel::add_var(Ca_er,"Ca_er",0,2.0,1.0);

        SimpleModel::add_param(K_R,"K_R",1,3,2);
        SimpleModel::add_param(k,"k",0,0.5,0.17);
        SimpleModel::add_param(K_2,"K_2",0.4,2,1);
        SimpleModel::add_param(v0,"v0",0.01,0.1,0.08);
        SimpleModel::add_param(v1,"v1",0.01,0.1,0.08);
        SimpleModel::add_param(beta,"beta",0,0.5,0.1);
        SimpleModel::add_param(K_A,"K_A",0.4,1,0.5);
        SimpleModel::add_param(rho_er,"rho_er",0.9,5,1);
        SimpleModel::add_param(VM2,"VM2",0,1.0,0.83);
        SimpleModel::add_param(VM3,"VM3",1,20,6.8);
        SimpleModel::add_param(k_f,"k_f",0,0.05,0.017);
    }

    ~AutoModel() {}

    vector<double> system(const vector<double>& v, const vector<double>& p)
    {
        vector<double> d(VAR_END);

        double Vin = p[v0] + p[v1] * p[beta];
        double Vout = p[k] * v[Ca_cyt];
        double Vserca = p[VM2] * hill(v[Ca_cyt],p[K_2],2);
        double Vrel = p[k_f] * v[Ca_er] + ( ( p[beta] * p[VM3] ) * hill(v[Ca_er],p[K_R],2) )
                                        * hill(v[Ca_cyt],p[K_A],4);

        d[Ca_cyt] = Vin - Vout + Vrel - Vserca;
        d[Ca_er] = p[rho_er] * ( Vserca - Vrel );

        return d;
    }
};

#endif /* ifndef AUTO_MODEL_H */
```

3

SYSTEM IDENTIFICATION PROGRAM

This Chapter gives Unified Modelling Language (UML) diagrams for the significant classes in the C++ program used to perform parameter estimation. The program evolved as new parameter estimation techniques were investigated. Nevertheless, the program was structured using pure virtual classes that define an interface that new added algorithms must follow. A simple example of such an interface is shown in Figure 54, where different integration algorithms can be used for generating a time series.

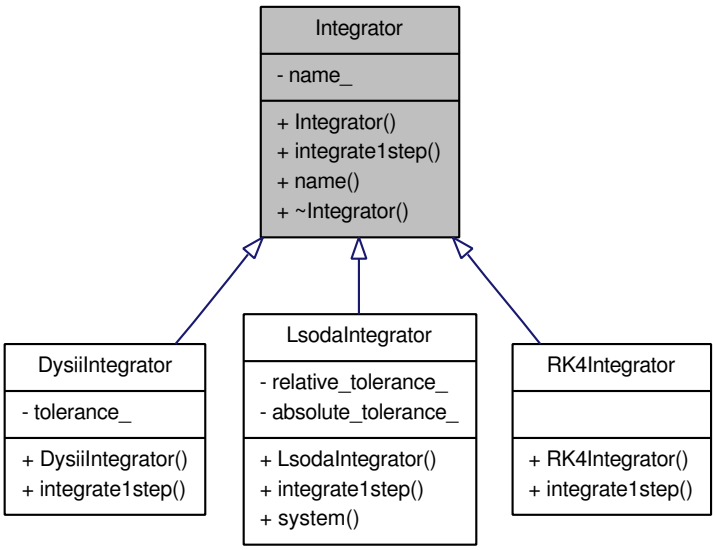


Figure 54: UML diagram of the integrators available for parameter estimation.

The type of integrator being used is specified in a configuration file following the inversion of control principle:

```
integrator:
{
    type = "rk4";
};
```

This configuration entry causes the creation of an integrator that uses a 4th order Runge Kutta algorithm. This integrator is then used by the rest of the program through the interface described by the pure virtual class `Integrator`. Following the same principles, the

algorithms used by the parameter estimator can be configured in a single file which can be saved with results.

An example configuration file, showing parameter estimation using multiple shooting, is shown below:

```
general:
{
   seed = 1;
   model = "lorenz";
   estimation_log = "sres_qnips_estimate.log";
   integration_log = "sres_qnips_integrate.log";
};

time_series:
{
   file = "input/lorenz.trace";
   field = 2;
   variable = "x1";
   sample_time = 0.02;
   noise = 2.0;
};

parameter_estimator:
{
   algorithm = "multiple_shooting";
   segment_size = 15;
   log_file = "sres_qnips_ms.log";

   optimiser:
   {
      algorithm = "hybrid";

      first_optimiser:
      {
         algorithm = "SRES";
         generations = 100;
         allow_parents = false;
         selected_parents = 30;
         children_from_selected = 200;
         using_constraints = false;
      };
```

```
      second_optimiser:
      {
         algorithm = "QNIPS";
         merit_function = "ArgaezTapia";
         output_file = "optpp.out";
         max_iterations = 150000;
         max_backtracks = 200;
         max_function_evaluations = 10000000;
      };
   };
};

integrator:
{
   type = "lsoda";
   relative_tolerance = 1e-2;
   absolute_tolerance = 0.0;
};
```

The main interfaces and classes that are created by the configuration files are shown in the UML diagrams that follow.

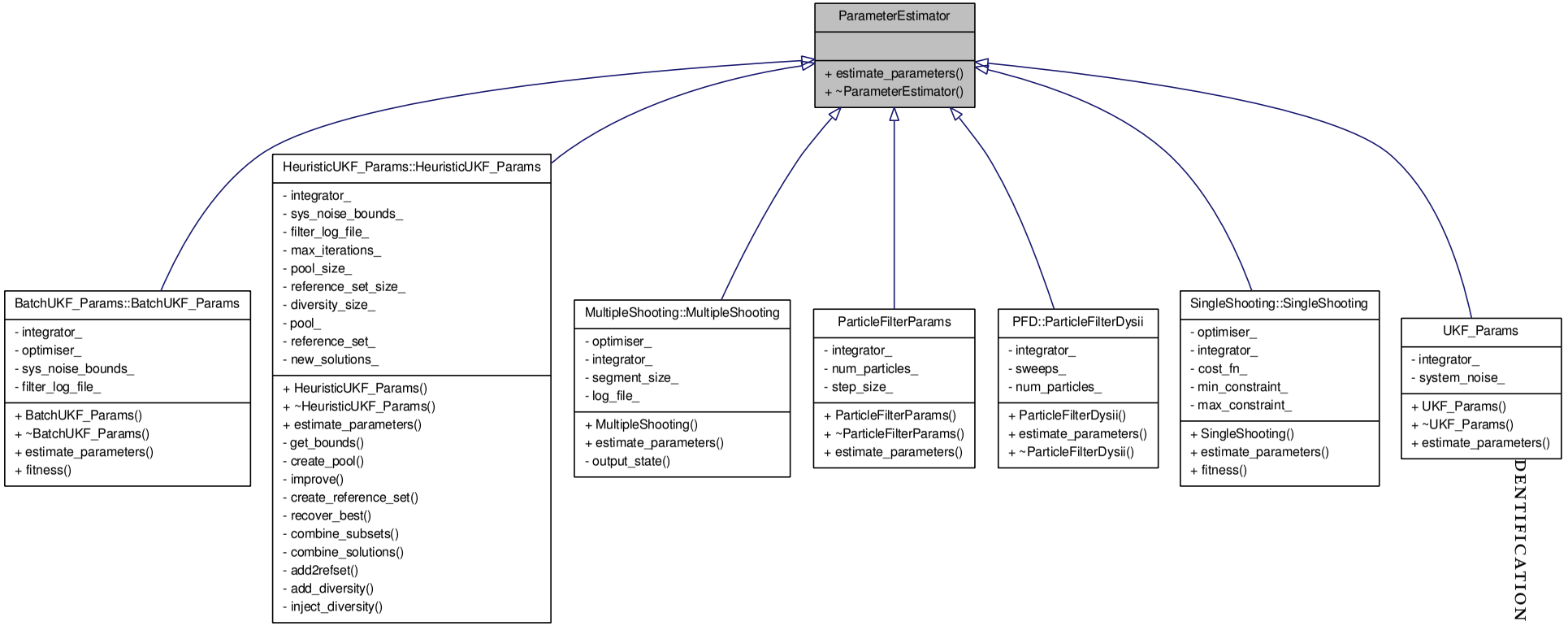


Figure 55: UML diagram of available parameter estimation algorithms.

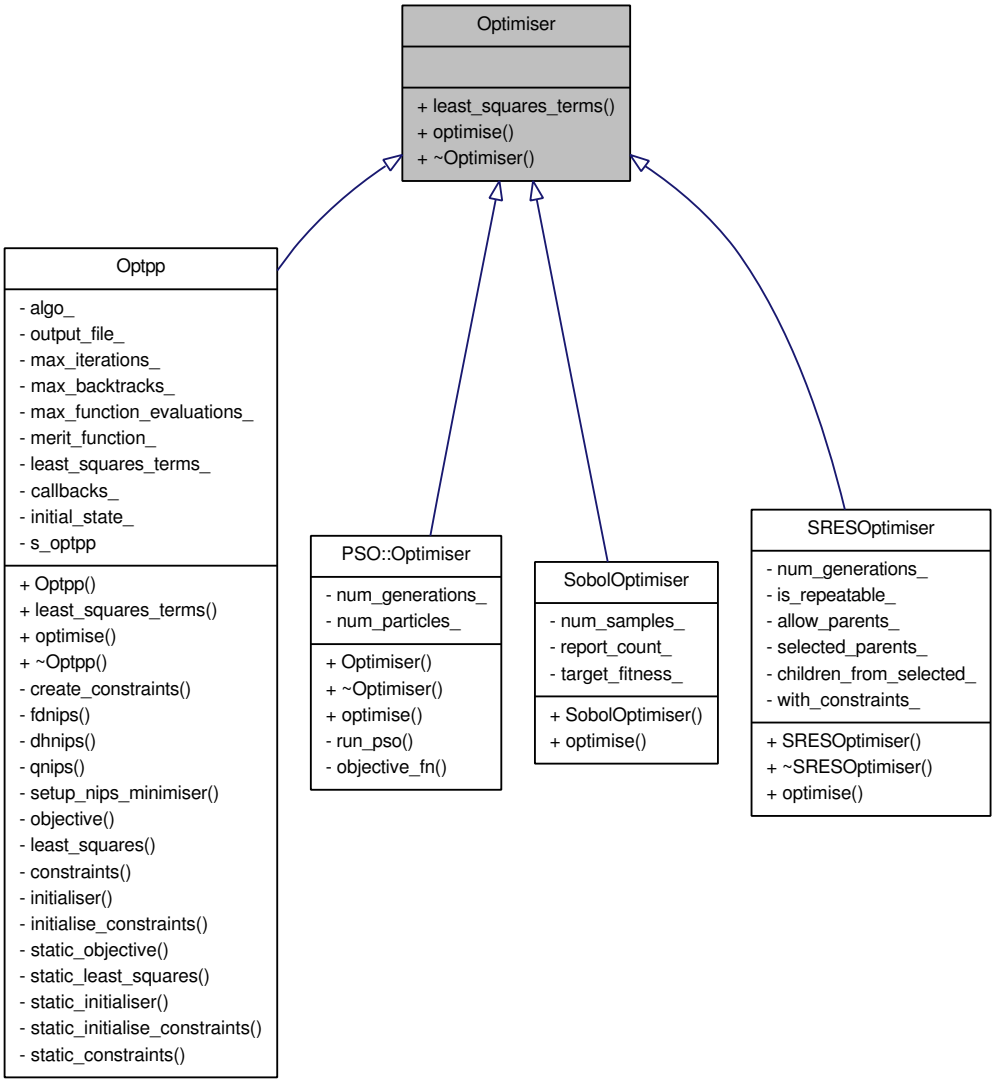


Figure 56: UML diagram of available optimisation algorithms.

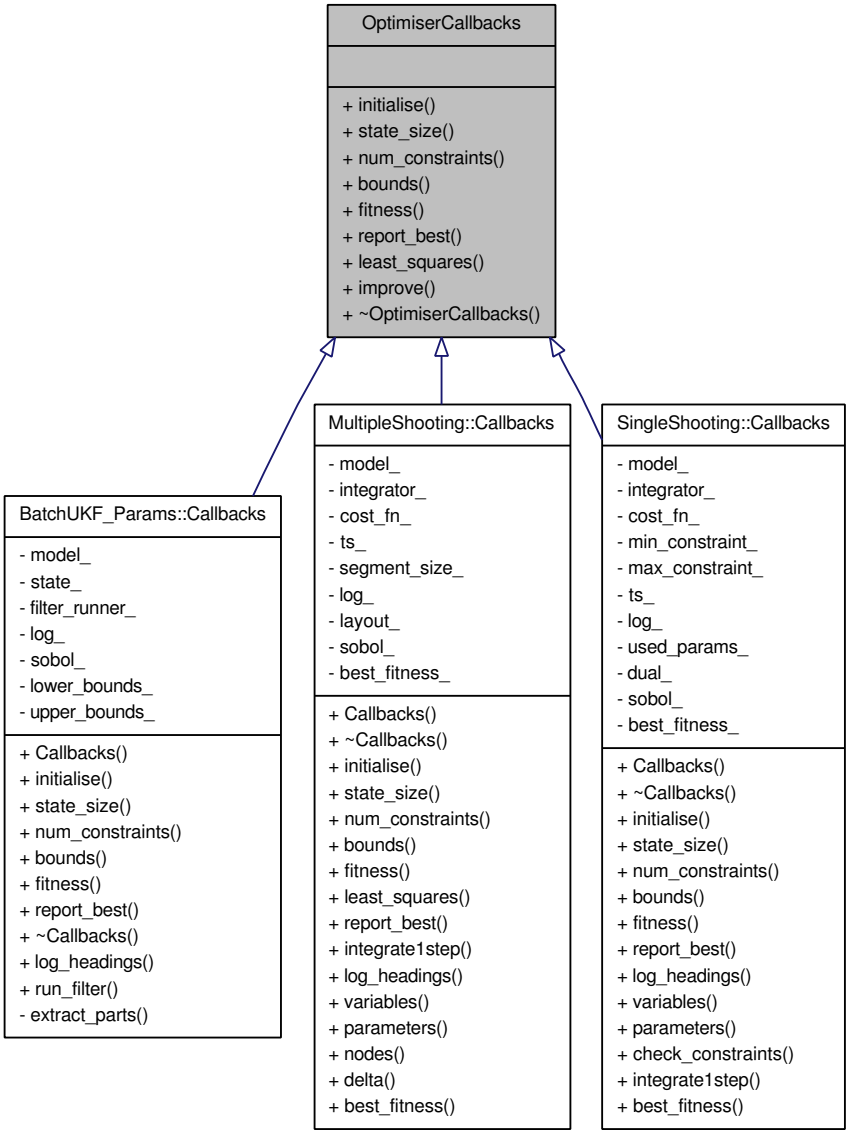


Figure 57: UML diagram of classes that connect an optimiser to the optimisation problem of performing parameter estimation.

GLOSSARY

ANGIOSPERM Flowering plant.

AUTOINHIBITORY DOMAIN A domain of a protein that can inhibit other domains of the same protein.

BURSTING Rapid nonperiodic Ca^{2+} oscillations.

CALMODULIN A ubiquitous Ca^{2+} binding protein.

CATION A positively charged ion.

CLADE A single common ancestor and all the decendents of that ancestor.

CORTICAL CELLS Cells that lie in a layer between the surface cells of a plant root and the conducting tissues further in.

ENDOCYTOSIS The entry of foreign bodies into the interior of a cell that remain isolated from the inside of the cell by cell wall material.

ENDOPLASMIC RETICULUM A membrane network within eukaryotic cells.

EPIDERMAL LAYER The outermost layer of cells in a plant root.

FLAVANOID A class of organic compound.

HABER-BOSCH PROCESS An industrial process for removing Nitrogen from the air and fixing it as ammonia.

LEGUME FAMILY A family of plants that mostly produce seed pods.

medicago truncatula A legume (common name Barrel Medic) used as a model plant for the study of legumes.

MYCELIAL The branched filaments of fungi.

NUCLEAR ENVELOPE The two membrane layers, and the space between them, that separate the nucleus from the cytosol.

NUCLEAR PORE Large protein complex that transports components through the nuclear envelope.

ODE Ordinary Differential Equation

ORGANOGENESIS The formation of tissues produced from undifferentiated cells that become differentiated.

PRIMORDIUM A tissue in its earliest stages of development.

SARCOPLASMIC RETICULUM A membrane network within muscle cells.

SRES Stochastic Ranking Evolutionary Strategy

UKF Unscented Kalman Filter

UML Unified Modeling Language

UV Ultraviolet

BIBLIOGRAPHY

[1] H. Akaike. A new look at the statistical model identification. *Automatic Control, IEEE Transactions on*, 19(6):716–723, 1974. (Cited on pages 26 and 79.)

[2] J.-M. Ane, G. B. Kiss, B. K. Riely, R. V. Penmetsa, G. E. D. Oldroyd, C. Ayax, J. Levy, F. Debelle, J.-M. Baek, P. Kalo, C. Rosenberg, B. A. Roe, S. R. Long, J. Denarie, and D. R. Cook. Medicago truncatula DMI1 Required for Bacterial and Fungal Symbioses in Legumes. *Science*, 303(5662):1364–1367, 2004. doi: 10.1126/science.1092986. URL `http://www.sciencemag.org/cgi/content/abstract/303/5662/1364`. (Cited on page 15.)

[3] T. Aparicio, E. F. Pozo, and D. Saura. Detecting determinism using recurrence quantification analysis: Three test procedures. *Journal of Economic Behavior & Organization*, 65(3-4):768–787, 2008. (Cited on pages 32 and 45.)

[4] E. Baake, M. Baake, H. G. Bock, and K. M. Briggs. Fitting ordinary differential equations to chaotic data. *Phys. Rev. A*, 45 (8):5524–5529, Apr 1992. doi: 10.1103/PhysRevA.45.5524. (Cited on pages 81 and 83.)

[5] V. Babovic and M. Keijzer. Genetic programming as a model induction engine. *Journal of Hydroinformatics*, 2(1):35–60, 2000. (Cited on page 77.)

[6] T. Bauchsbaum and S. Voessner. *Genetic Programming*, chapter Information-Dependent Switching of Identification Criteria in a Genetic Programming System for System Identification, pages 300–309. Springer Berlin / Heidelberg, 2006. (Cited on page 84.)

[7] L. Becks, F. M. Hilker, H. Malchow, K. Jurgens, and H. Arndt. Experimental demonstration of chaos in a microbial food web. *Nature*, 435(7046):1226–1229, 2005. (Cited on page 108.)

[8] M. J. Berridge, P. Lipp, and M. D. Bootman. The versatility and universality of calcium signalling. *Nat Rev Mol Cell Biol*, 1 (1):11–21, 10 2000. URL `http://dx.doi.org/10.1038/35036035`. (Cited on page 18.)

[9] R. Bertram and A. Sherman. A calcium-based phantom bursting model for pancreatic islets. *Bulletin of Mathematical Biology*, 66(5): 1313–1344, 09 2004. URL `http://dx.doi.org/10.1016/j.bulm.2003.12.005`. (Cited on page 85.)

[10] H.-G. Beyer and H.-P. Schwefel. Evolution strategies –a comprehensive introduction. *Natural Computing: an international journal*, 1(1):3–52, 2002. ISSN 1567-7818. doi: http://dx.doi.org/10.1023/A:1015059928466. (Cited on page 55.)

[11] H. Bien, L. Yin, and E. Entcheva. Calcium instabilities in mammalian cardiomyocyte networks. *Biophys. J.*, 90(7):2628–2640, 2006. (Cited on page 45.)

[12] H. G. Bock, E. Kostina, and J. P. Schlöder. Numerical methods for parameter estimation in nonlinear differential algebraic equations. *GAMM-Mitteilungen*, 30(2):376–408, 2007. (Cited on page 81.)

[13] J. M. Borghans, G. Dupont, and A. Goldbeter. Complex intracellular calcium oscillations a theoretical exploration of possible mechanisms. *Biophys Chem*, 66(1):25–41, 1997. ISSN 0301-4622 (Print). (Cited on page 23.)

[14] W. Bridewell, P. Langley, L. Todorovski, and S. Džeroski. Inductive process modeling. *Machine Learning*, 71(1):1–32, 04 2008. (Cited on pages 77 and 78.)

[15] C. Brière, T. C. Xiong, C. Mazars, and R. Ranjeva. Autonomous regulation of free Ca^{2+} concentrations in isolated plant cell nuclei: a mathematical analysis. *Cell Calcium*, 39(4):293–303, Apr 2006. doi: 10.1016/j.ceca.2005.11.005. (Cited on pages 89 and 106.)

[16] P. J. Brockwell and R. A. Davis. *Introduction to Time Series and Forecasting*. Springer-Verlag, second edition, 2002. (Cited on pages 26 and 28.)

[17] T. Capiod, J. Noel, L. Combettes, and M. Claret. Cyclic AMP-evoked oscillations of intracellular [Ca^{2+}] in guinea-pig hepatocytes. *Biochem J*, 275 (Pt 1)(0264-6021 (Print)):277–80, 1991. (Cited on page 23.)

[18] M. Charpentier, R. Bredemeier, G. Wanner, N. Takeda, E. Schleiff, and M. Parniske. Lotus japonicus CASTOR and POLLUX Are Ion Channels Essential for Perinuclear Calcium Spiking

in Legume Root Endosymbiosis. *Plant Cell*, 20(12):3467–3479, 2008. doi: 10.1105/tpc.108.063255. URL http://www.plantcell.org/cgi/content/abstract/20/12/3467. (Cited on page 18.)

[19] Y. Chen. Initializing a hurricane vortex with an ensemble kalman filter. In *27th Conference on Hurricanes and Tropical Meteorology*, 2006. (Cited on page 83.)

[20] C. J. Dixon, N. M. Woods, K. S. Cuthbertson, and P. H. Cobbold. Evidence for two Ca^{2+}-mobilizing purinoceptors on rat hepatocytes. *Biochem. J.*, 269(2):499–502, 1990. (Cited on page 23.)

[21] G. Dupont and A. Goldbeter. One-pool model for Ca^{2+} oscillations involving Ca^{2+} and inositol 1,4,5-trisphosphate as co-agonists for Ca^{2+} release. *Cell Calcium*, 14(4):311–22, 1993. ISSN 0143-4160 (Print). (Cited on pages 28, 78, 79, and 107.)

[22] J.-P. Eckmann, S. Oliffson Kamphorst, and D. Ruelle. Recurrence plots of dynamical systems. *Europhysics Letters*, 4:973–+, Nov. 1987. (Cited on page 31.)

[23] A. Edwards, A. B. Heckmann, F. Yousafzai, G. Duc, and J. A. Downie. Structural implications of mutations in the pea SYM8 symbiosis gene, the DMI1 ortholog, encoding a predicted ion channel. *Molecular Plant-Microbe Interactions*, 20(10):1183–1191, 2007. doi: 10.1094/MPMI-20-10-1183. URL http://apsjournals.apsnet.org/doi/abs/10.1094/MPMI-20-10-1183. PMID: 17918620. (Cited on page 86.)

[24] D. W. Ehrhardt, R. Wais, and S. R. Long. Calcium spiking in plant root hairs responding to rhizobium nodulation signals. *Cell*, 85 (5):673–81, 1996. ISSN 0092-8674 (Print). (Cited on pages 16, 17, and 18.)

[25] S. Elner, D. Nychka, and A. Gallant. Lenns, a program to estimate the dominant lyapunov exponent of noisy nonlinear systems from time series data. *Institute of Statistics Mimeo Series, Statistics Department, North Carolina State University*, (2235), 1992. (Cited on page 35.)

[26] B. Ermentrout. *Simulating, Analyzing, and Animating Dynamical Systems: A Guide to XPPAUT for Researchers and Students*. Software, Environments, and Tools. SIAM. (Cited on page 105.)

[27] G. Evensen. The ensemble kalman filter: theoretical formulation and practical implementation. *Ocean Dynamics*, 53(4):343–367, 2003. (Cited on page 83.)

[28] M. Falcke. Reading the patterns in living cells — the physics of Ca^{2+} signaling. *Advances in Physics*, 53:255–440, May 2004. (Cited on page 18.)

[29] M. Falcke. On the role of stochastic channel behavior in intracellular Ca^{2+} dynamics. *Biophys. J.*, 84(1):42–56, 2003. (Cited on page 23.)

[30] D. J. Gage. Infection and invasion of roots by symbiotic, nitrogen-fixing rhizobia during nodulation of temperate legumes. *Microbiol Mol Biol Rev*, 68(2):280–300, 2004. ISSN 1092-2172 (Print). (Cited on page 12.)

[31] J. N. Galloway and E. B. Cowling. Reactive nitrogen and the world: 200 years of change. *Ambio*, 31(2):64–71, 2002. ISSN 0044-7447 (Print). (Cited on page 11.)

[32] P. Gaspard. *Chaos, Scattering and Statistical Mechanics*. Nonlinear Science Series. Cambridge University Press, 2005. (Cited on page 45.)

[33] H. Gherbi, K. Markmann, S. Svistoonoff, J. Estevan, D. Autran, G. Giczey, F. Auguy, B. Péret, L. Laplaze, C. Franche, M. Parniske, and D. Bogusz. SymRK defines a common genetic basis for plant root endosymbioses with arbuscular mycorrhiza fungi, rhizobia, and Frankiabacteria. *Proceedings of the National Academy of Sciences*, 105(12):4928–4932, 2008. doi: 10.1073/pnas.0710618105. URL `http://www.pnas.org/content/105/12/4928.abstract`. (Cited on page 15.)

[34] W. R. Gilks and C. Berzuini. Following a moving target — monte carlo inference for dynamic bayesian models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63:127–146(20), 2001. (Cited on pages 68 and 83.)

[35] C. Gleason, S. Chaudhuri, T. Yang, A. Munoz, B. W. Poovaiah, and G. E. D. Oldroyd. Nodulation independent of rhizobia induced by a calcium-activated kinase lacking autoinhibition. *Nature*, 441(7097):1149–52, 2006. ISSN 1476-4687 (Electronic). (Cited on page 15.)

[36] A. K. Green, C. J. Dixon, A. G. McLennan, P. H. Cobbold, and M. J. Fisher. Adenine dinucleotide-mediated cytosolic free Ca^{2+} oscillations in single hepatocytes. *FEBS Letters*, 322(2):197–200, 1993. (Cited on page 23.)

[37] N. Greenwood and A. Earnshaw. *Chemistry of the elements*. Elsevier, second edition, 1997. (Cited on page 21.)

[38] C. Grygorczyk and R. Grygorczyk. A Ca^{2+} and voltage-dependent cation channel in the nuclear envelope of red beet. *Biochimica et Biophysica Acta (BBA) - Biomembranes*, 1375(1-2):117 – 130, 1998. ISSN 0005-2736. (Cited on pages 85 and 89.)

[39] T. Haberichter, M. Marhl, and R. Heinrich. Birhythmicity, trirhythmicity and chaos in bursting calcium oscillations. *Biophysical Chemistry*, 90(1):17–30, 2001. (Cited on pages 23, 26, 45, and 80.)

[40] M. H. Hansen and B. Yu. Model selection and the principle of minimum description length. *Journal of the American Statistical Association*, 96(454):746–774, June 2001. (Cited on page 83.)

[41] S. Hayes, C. Grebogi, E. Ott, and A. Mark. Experimental control of chaos for communication. *Phys. Rev. Lett.*, 73(13):1781–1784, Sep 1994. (Cited on page 48.)

[42] S. Hazledine, J. Sun, D. Wysham, J. A. Downie, G. E. D. Oldroyd, and R. J. Morris. Nonlinear time series analysis of nodulation factor induced calcium oscillations: Evidence for deterministic chaos? *PLoS ONE*, 4(8):e6637, 08 2009. doi: 10.1371/journal.pone.0006637. (Cited on page 20.)

[43] Q. He, L. Wang, and B. Liu. Parameter estimation for chaotic systems by particle swarm optimization. *Chaos, Solitons & Fractals*, 34(2):654–661, 2007. (Cited on pages 55, 57, 66, and 83.)

[44] R. Hegger and H. Kantz. Improved false nearest neighbor method to detect determinism in time series data. *Phys. Rev. E*, 60(4):4970–4973, Oct 1999. (Cited on page 29.)

[45] N. E. Huang, Z. Shen, S. R. Long, M. C. Wu, H. H. Shih, Q. Zheng, N. C. Yen, C. C. Tung, and H. H. Liu. The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 454 (1971):903–995, 1998. (Cited on page 28.)

[46] B. A. Hungate, J. S. Dukes, M. R. Shaw, Y. Luo, and C. B. Field. Nitrogen and Climate Change. *Science*, 302(5650):1512–1513, 2003. (Cited on page 11.)

[47] H. Iba. Inference of differential equation models by genetic programming. *Inf. Sci.*, 178(23):4453–4468, 2008. ISSN 0020-0255. doi: http://dx.doi.org/10.1016/j.ins.2008.07.029. (Cited on page 50.)

[48] Invitrogen. Molecular probes the handbook, January 2010. URL `http://www.invitrogen.com/site/us/en/home/References/Molecular-Probes-The-Handbook.html`. (Cited on page 16.)

[49] M. Isard and A. Blake. Condensation —conditional density propagation forvisual tracking. *Int. J. Comput. Vision*, 29(1):5–28, 1998. ISSN 0920-5691. (Cited on pages 68 and 83.)

[50] E. M. Izhikevich. *Dynamical Systems in Neuroscience*. The Mit Press, 1 edition, November 2006. (Cited on page 87.)

[51] E. S. Jensen and H. Hauggaard-Nielsen. How can increased use of biological N_2 fixation in agriculture benefit the environment? *Plant and Soil*, 252(1):177–186, 2003. (Cited on page 11.)

[52] X. Ji and Y. Xu. libSRES: a C library for stochastic ranking evolution strategy for parameter estimation. *Bioinformatics*, 22 (1):124–126, 2006. (Cited on page 84.)

[53] E.-P. Journet, N. El-Gachtouli, V. Vernoud, F. de Billy, M. Pichon, A. Dedieu, C. Arnould, D. Morandi, D. G. Barker, and V. Gianinazzi-Pearson. Medicago truncatula ENOD11: A novel rprp-encoding early nodulin gene expressed during mycorrhization in arbuscule-containing cells. *Molecular Plant-Microbe Interactions*, 14(6):737–748, 2001. doi: 10.1094/MPMI.2001.14.6.737. URL `http://apsjournals.apsnet.org/doi/abs/10.1094/MPMI.2001.14.6.737`. PMID: 11386369. (Cited on pages 15 and 17.)

[54] K. Judd. Chaotic-time-series reconstruction by the bayesian paradigm: Right results by wrong methods. *Phys. Rev. E*, 67(2): 026212, Feb 2003. (Cited on page 57.)

[55] S. Julier and J. Uhlmann. A new extension of the kalman filter to nonlinear systems, 1997. (Cited on pages 63, 65, and 83.)

[56] H. Kantz and T. Schreiber. *Nonlinear time series analysis*. Cambridge University Press, New York, NY, USA, 1997. ISBN 0-521-55144-7. (Cited on pages 29 and 33.)

[57] D. T. Kaplan and L. Glass. Coarse-grained embeddings of time series: random walks, gaussian random processes, and deterministic chaos. *Physica D Nonlinear Phenomena*, 64:431–454, Apr. 1993. (Cited on page 32.)

[58] D. T. Kaplan and L. Glass. Direct test for determinism in a time series. *Phys. Rev. Lett.*, 68(4):427–430, Jan 1992. (Cited on page 32.)

[59] M. Kejzer and V. Babovic. Dimensionally aware genetic programming. In W. Banzhaf, J. Daida, A. E. Eiben, M. H. Garzon, V. Honavar, M. Jakiela, and R. E. Smith, editors, *Proceedings of the Genetic and Evolutionary Computation Conference*, volume 2, pages 1069–1076, 1999. (Cited on page 77.)

[60] M. B. Kennel. Statistical test for dynamical nonstationarity in observed time-series data. *Phys. Rev. E*, 56(1):316–321, Jul 1997. (Cited on page 31.)

[61] S. Kikuchi, D. Tominaga, M. Arita, K. Takahashi, and M. Tomita. Dynamic modeling of genetic networks using genetic algorithm and S-system. *Bioinformatics*, 19(5):643–650, 2003. (Cited on page 50.)

[62] C. Kistner and M. Parniske. Evolution of signal transduction in intracellular symbiosis. *Trends in Plant Science*, 7(11):511–518, 2002. (Cited on pages 11 and 15.)

[63] C. B. Klee, T. H. Crouch, and P. G. Richman. Calmodulin. *Annual Review of Biochemistry*, 49(1):489–515, 1980. doi: 10.1146/annurev.bi.49.070180.002421. URL `http://arjournals.annualreviews.org/doi/abs/10.1146/annurev.bi.49.070180.002421`. (Cited on pages 98 and 99.)

[64] M. Korenberg. Identifying nonlinear difference equation and functional expansion representations: The fast orthogonal algorithm. *Annals of Biomedical Engineering*, 16(1):123–142, 1988. (Cited on page 50.)

[65] S. Kosuta, S. Hazledine, J. Sun, H. Miwa, R. J. Morris, J. A. Downie, and G. E. D. Oldroyd. Differential and chaotic calcium

signatures in the symbiosis signaling pathway of legumes. *Proceedings of the National Academy of Sciences*, 105(28):9823–9828, 2008. (Cited on pages 15, 17, 21, and 48.)

[66] D. Kugiumtzis. State space reconstruction parameters in the analysis of chaotic time series - the role of the time window length. *Physica D*, 95:13–28, 1996. (Cited on page 29.)

[67] U. Kummer, L. F. Olsen, C. J. Dixon, A. K. Green, E. Bornberg-Bauer, and G. Baier. Switching from simple to complex oscillations in calcium signaling. *Biophys J*, 79(3):1188–95, 2000. ISSN 0006-3495 (Print). (Cited on page 23.)

[68] U. Kummer, B. Krajnc, J. Pahle, A. K. Green, C. J. Dixon, and M. Marhl. Transition from Stochastic to Deterministic Behavior in Calcium Oscillations. *Biophys. J.*, 89(3):1603–1611, 2005. (Cited on page 23.)

[69] Y.-C. Lai. Persistence of supertransients of spatiotemporal chaotic dynamical systems in noisy environment. *Physics Letters A*, 200(6):418 – 422, 1995. ISSN 0375-9601. (Cited on page 47.)

[70] P. Langley, O. Shiran, J. Shrager, L. Todorovski, and A. Pohorille. Constructing explanatory process models from biological data and knowledge. *Artificial Intelligence in Medicine*, 37(3):191–201, 2006. (Cited on page 77.)

[71] J. Levy, C. Bres, R. Geurts, B. Chalhoub, O. Kulikova, G. Duc, E.-P. Journet, J.-M. Ane, E. Lauber, T. Bisseling, J. Denarie, C. Rosenberg, and F. Debelle. A Putative Ca^{2+} and Calmodulin-Dependent Protein Kinase Required for Bacterial and Fungal Symbioses. *Science*, 303(5662):1361–1364, 2004. doi: 10.1126/science.1093038. URL `http://www.sciencemag.org/cgi/content/abstract/303/5662/1361`. (Cited on page 15.)

[72] F. G. P. Lhuissier, N. C. A. De Ruijter, B. J. Sieberer, J. J. Esseling, and A. M. C. Emons. Time course of cell biological events evoked in legume root hairs by rhizobium nod factors: State of the art. *Ann Bot*, 87(3):289–302, 2001. (Cited on page 12.)

[73] Y.-X. Li and J. Rinzel. Equations for InsP3 receptor-mediated Ca^{2+} oscillations derived from a detailed kinetic model: A hodgkin-huxley like formalism. *Journal of Theoretical Biology*, 166(4):461–473, 1994. (Cited on page 79.)

[74] L. Ljung. *System Identification Theory for the User*. Prentice Hall, second edition, 1999. (Cited on pages 49 and 50.)

[75] J. C. W. Locke, A. J. Millar, and M. S. Turner. Modelling genetic networks with noisy and varied experimental data: the circadian clock in arabidopsis thaliana. *J Theor Biol*, 234(3):383–393, 2005. ISSN 0022-5193 (Print). (Cited on pages 52 and 54.)

[76] E. N. Lorenz. Deterministic nonperiodic flow. *Journal of Atmospheric Sciences*, 20:130–141, 1963. (Cited on pages 26 and 53.)

[77] M. Marhl, S. Schuster, M. Brumen, and R. Heinrich. Modelling the interrelations between calcium oscillations and ER membrane potential oscillations. *Biophysical Chemistry*, 63(2-3):221–239, 1997. (Cited on pages 79, 98, and 104.)

[78] K. Markmann, G. Giczey, and M. Parniske. Functional adaptation of a plant receptor- kinase paved the way for the evolution of intracellular root symbioses with bacteria. *PLoS Biol*, 6(3):e68, 03 2008. doi: 10.1371/journal.pbio.0060068. (Cited on page 15.)

[79] M. Mazzanti, J. O. Bustamante, and H. Oberleithner. Electrical Dimension of the Nuclear Envelope. *Physiol. Rev.*, 81(1):1–19, 2001. URL `http://physrev.physiology.org/cgi/content/abstract/81/1/1`. (Cited on page 105.)

[80] P. E. McSharry and L. A. Smith. Better nonlinear models from noisy data: Attractors with maximum likelihood. *Phys. Rev. Lett.*, 83(21):4285–4288, Nov 1999. (Cited on page 54.)

[81] R. V. D. Merwe and E. A. Wan. The square-root unscented kalman filter for state and parameter-estimation. In *in International Conference on Acoustics, Speech, and Signal Processing*, pages 3461–3464, 2001. (Cited on page 63.)

[82] H. Miwa, J. Sun, G. E. D. Oldroyd, and J. A. Downie. Analysis of calcium spiking using a cameleon calcium sensor reveals that nodulation gene expression is regulated by calcium spike number and the developmental status of the cell. *Plant J*, 48 (6):883–894, 2006. ISSN 0960-7412 (Print). (Cited on pages 16 and 17.)

[83] H. Miwa, J. Sun, G. E. D. Oldroyd, and J. A. Downie. Analysis of nod-factor-induced calcium signaling in root hairs of symbiotically defective mutants of lotus japonicus. *Mol Plant Microbe*

Interact, 19(8):914–23, 2006. ISSN 0894-0282 (Print). (Cited on page 17.)

[84] C. G. Moles, P. Mendes, and J. R. Banga. Parameter estimation in biochemical pathways: A comparison of global optimization methods. *Genome Research*, 13(11):2467–2474, 2003. (Cited on pages 54 and 57.)

[85] D. N. Mukhin, A. M. Feigin, E. M. Loskutov, and Y. I. Molkov. Modified bayesian approach for the reconstruction of dynamical systems from time series. *Physical Review E (Statistical, Nonlinear, and Soft Matter Physics)*, 73(3):036211, 2006. (Cited on page 81.)

[86] T. Nagata, S. Iizumi, K. Satoh, H. Ooka, J. Kawai, P. Carninci, Y. Hayashizaki, Y. Otomo, K. Murakami, K. Matsubara, and S. Kikuchi. Comparative analysis of plant and animal calcium signal transduction element using plant full-length cDNA data. *Mol Biol Evol*, 21(10):1855–70, 2004. ISSN 0737-4038 (Print). (Cited on pages 18 and 24.)

[87] D. Napoletani and T. D. Sauer. Reconstructing the topology of sparsely connected dynamical networks. *Physical Review E*, 77(2): 026103–+, Feb. 2008. doi: 10.1103/PhysRevE.77.026103. (Cited on page 50.)

[88] D. Napoletani, T. Sauer, D. C. Struppa, E. Petricoin, and L. Liotta. Sparse Dynamical Network Reconstruction: the EGFR network case. *ArXiv e-prints*, May 2007. (Cited on page 50.)

[89] J. Nocedal and S. Wright. *Numerical Optimization*. Springer Series in Operations Research and Financial Engineering. Springer Science, 2 edition, 2006. (Cited on pages 81 and 83.)

[90] N. Noman and H. Iba. Inference of gene regulatory networks using s-system and differential evolution. In *GECCO '05: Proceedings of the 2005 conference on Genetic and evolutionary computation*, pages 439–446, New York, NY, USA, 2005. ACM. ISBN 1-59593-010-8. doi: http://doi.acm.org/10.1145/1068009.1068079. (Cited on page 50.)

[91] G. E. D. Oldroyd and J. A. Downie. Calcium, kinases and nodulation signalling in legumes. *Nat Rev Mol Cell Biol*, 5(7): 566–76, 2004. ISSN 1471-0072 (Print). (Cited on pages 12 and 21.)

[92] G. E. D. Oldroyd and J. A. Downie. Nuclear calcium changes at the core of symbiosis signalling. *Curr Opin Plant Biol*, 9(4):351–7, 2006. ISSN 1369-5266 (Print). (Cited on pages 12 and 24.)

[93] E. Ott, C. Grebogi, and J. A. Yorke. Controlling chaos. *Phys. Rev. Lett.*, 64(11):1196–1199, Mar 1990. (Cited on page 48.)

[94] D. J. D. Pauw and B. D. Baets. Incorporating model identifiability into equation discovery of ODE systems. In *GECCO '08: Proceedings of the 2008 GECCO conference companion on Genetic and evolutionary computation*, pages 2135–2140, New York, NY, USA, 2008. ACM. ISBN 978-1-60558-131-6. (Cited on page 84.)

[95] E. Peiter, J. Sun, A. B. Heckmann, M. Venkateshwaran, B. K. Riely, M. S. Otegui, A. Edwards, G. Freshour, M. G. Hahn, D. R. Cook, D. Sanders, G. E. Oldroyd, J. A. Downie, and J.-M. Ane. The medicago truncatula DMI1 protein modulates cytosolic calcium signaling. *Plant Physiol.*, 145(1):192–203, 2007. (Cited on page 15.)

[96] M. Perc and M. Marhl. Sensitivity and flexibility of regular and chaotic calcium oscillations. *Biophysical Chemistry*, 104(2): 509–522, 2003. (Cited on page 47.)

[97] M. Perc, A. K. Green, C. J. Dixon, and M. Marhl. Establishing the stochastic nature of intracellular calcium oscillations from experimental data. *Biophysical Chemistry*, 132(1):33–38, 2008. (Cited on pages 23, 45, and 46.)

[98] O. H. Petersen, O. V. Gerasimenko, J. V. Gerasimenko, H. Mogami, and A. V. Tepikin. The calcium store in the nuclear envelope. *Cell Calcium*, 23(2-3):87 – 90, 1998. ISSN 0143-4160. Nuclear calcium: regulation and functions. (Cited on pages 18 and 89.)

[99] L. Petzold. Automatic selection of methods for solving stiff and nonstiff systems of ordinary differential equations. *SIAM Journal on Scientific and Statistical Computing*, 4(1):136–148, 1983. doi: 10.1137/0904010. URL `http://link.aip.org/link/?SCE/4/136/1`. (Cited on page 84.)

[100] R. Poli, W. B. Langdon, and N. F. McPhee. *A Field Guide to Genetic Programming*. http://lulu.com, 2008. (Cited on page 70.)

[101] C.-S. Poon and M. Barahona. Titration of chaos with added noise. *Proceedings of the National Academy of Science*, 98:7107–7112, June 2001. (Cited on pages 33 and 46.)

[102] C.-S. Poon and C. K. Merrill. Decrease of cardiac chaos in congestive heart failure. *Nature*, 389(6650):492–495, 1997. (Cited on page 45.)

[103] W. Press, S. Teukolsky, W. Vetterling, and B. Flannery. *Numerical Recipes*. Cambridge University Press, Cambridge, UK, 3rd edition, 2007. (Cited on page 57.)

[104] B. K. Riely, J.-H. Mun, and J.-M. Ane. Unravelling the molecular basis for symbiotic signal transduction in legumes. *Molecular Plant Pathology*, 7(3):197–207, 2006. (Cited on page 12.)

[105] B. K. Riely, G. Lougnon, J.-M. Anè, and D. R. Cook. The symbiotic ion channel homolog DMI1 is localized in the nuclear membrane of *Medicago truncatula* roots. *The Plant Journal*, 49(2): 208–216, 2007. (Cited on page 18.)

[106] M. Rodriguez-Fernandez, P. Mendes, and J. R. Banga. A hybrid approach for efficient and robust parameter estimation in biochemical pathways. *Biosystems*, 83(2-3):248–265, 2006. ISSN 0303-2647 (Print). (Cited on pages 55 and 81.)

[107] M. T. Rosenstein, J. J. Collins, and C. J. D. Luca. A practical method for calculating largest lyapunov exponents from small data sets. *Physica D*, 65:117–134, 1993. (Cited on pages 29 and 34.)

[108] T. P. Runarsson and X. Yao. Stochastic ranking for constrained evolutionary optimization. *IEEE Transactions on Evolutionary Computation*, 4(3):284–294, 2000. (Cited on pages 57, 81, and 83.)

[109] E. Sakamoto and H. Iba. Inferring a system of differential equations for a gene regulatory network by using genetic programming. In *Proceedings of the 2001 Congress on Evolutionary Computation CEC2001*, pages 720–726, COEX, World Trade Center, 159 Samseong-dong, Gangnam-gu, Seoul, Korea, 27-30 2001. IEEE Press. ISBN 0-7803-6658-1. (Cited on pages 50 and 51.)

[110] T. Sauer, J. A. Yorke, and M. Casdagli. Embedology. *Journal of Statistical Physics*, V65(3):579–616, 1991. (Cited on page 29.)

[111] M. Schmidt and H. Lipson. Distilling free-form natural laws from experimental data. *Science*, 324(5923):81–85, 2009. (Cited on page 50.)

[112] T. Schreiber and A. Schmitz. Improved surrogate data for nonlinearity tests. *Phys. Rev. Lett.*, 77(4):635–638, Jul 1996. (Cited on pages 26, 27, and 33.)

[113] S. Schuster, M. Marhl, and T. Hofer. Modelling of simple and complex calcium oscillations. From single-cell responses to intercellular signalling. *Eur J Biochem*, 269(5):1333–55, 2002. ISSN 0014-2956 (Print). (Cited on pages 18, 23, 77, 78, and 79.)

[114] G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978. (Cited on page 83.)

[115] S. L. Shaw and S. R. Long. Nod factor elicits two separable calcium responses in medicago truncatula root hair cells. *Plant Physiol*, 131(3):976–84, 2003. ISSN 0032-0889 (Print). (Cited on page 17.)

[116] T. Shinbrot, C. Grebogi, J. A. Yorke, and E. Ott. Using small perturbations to control chaos. *Nature*, 363(6428):411–417, 1993. (Cited on page 48.)

[117] B. J. Sieberer, M. Chabaud, A. C. Timmers, A. Monin, J. Fournier, and D. G. Barker. A nuclear-targeted cameleon demonstrates intranuclear Ca^{2+} spiking in medicago truncatula root hairs in response to rhizobial nodulation factors. *Plant Physiol.*, 151(3): 1197–1206, 2009. doi: 10.1104/pp.109.142851. URL `http://www.plantphysiol.org/cgi/content/abstract/151/3/1197`. (Cited on pages 16, 102, and 105.)

[118] A. Sitz, U. Schwarz, J. Kurths, and H. U. Voss. Estimation of parameters and unobserved components for nonlinear systems from noisy time series. *Phys. Rev. E*, 66(1):016210, Jul 2002. (Cited on pages 61, 63, and 65.)

[119] A. Skupin, H. Kettenmann, U. Winkler, M. Wartenberg, H. Sauer, S. C. Tovey, C. W. Taylor, and M. Falcke. How does intracellular Ca^{2+} oscillate: By chance or by the clock? *Biophys. J.*, 94(6): 2404–2411, 2008. (Cited on pages 23, 26, and 40.)

[120] J. Sneyd, K. Tsaneva-Atanasova, D. I. Yule, J. L. Thompson, and T. J. Shuttleworth. Control of calcium oscillations by membrane fluxes. *Proceedings of the National Academy of Sciences of the United States of America*, 101(5):1392–1396, 2004. (Cited on page 48.)

[121] J. Sun, V. Cardoza, D. M. Mitchell, L. Bright, G. Oldroyd, and J. M. Harris. Crosstalk between jasmonic acid, ethylene and

nod factor signaling allows integration of diverse inputs for regulation of nodulation. *The Plant Journal*, 46(6):961–970, 2006. (Cited on page 26.)

[122] A. Szewczyk. The intracellular potassium and chloride channels: Properties, pharmacology and function (review). *Molecular Membrane Biology*, 15(2):49–58, 1998. URL http://www.informaworld.com/10.3109/09687689809027518. (Cited on page 104.)

[123] D. Thomas, S. C. Tovey, T. J. Collins, M. D. Bootman, M. J. Berridge, and P. Lipp. A comparison of fluorescent Ca^{2+} indicator properties and their use in measuring elementary and global Ca^{2+} signals. *Cell Calcium*, 28(4):213 – 223, 2000. ISSN 0143-4160. (Cited on page 99.)

[124] H. U. Voss, J. Timmer, and J. Kurths. Nonlinear dynamical system identification from uncertain and indirect measurements. *I. J. Bifurcation and Chaos*, 14(6):1905–1933, 2004. (Cited on pages 63 and 65.)

[125] R. J. Wais, C. Galera, G. Oldroyd, R. Catoira, R. V. Penmetsa, D. Cook, C. Gough, J. Denarie, and S. R. Long. Genetic analysis of calcium spiking responses in nodulation mutants of medicago truncatula. *Proc Natl Acad Sci U S A*, 97(24):13407–12, 2000. ISSN 0027-8424 (Print). (Cited on page 15.)

[126] R. J. Wais, D. H. Wells, and S. R. Long. Analysis of differences between sinorhizobium meliloti 1021 and 2011 strains using the host calcium spiking response. *Molecular Plant-Microbe Interactions*, 15(12):1245–1252, 2002. (Cited on page 24.)

[127] S. A. Walker, V. Viprey, and J. A. Downie. Dissection of nodulation signaling using pea mutants defective for calcium spiking induced by nod factors and chitin oligomers. *Proc Natl Acad Sci U S A*, 97(24):13413–8, 2000. ISSN 0027-8424 (Print). (Cited on page 16.)

[128] E. Wan and R. van der Merwe. The unscented kalman filter for nonlinear estimation, 2000. (Cited on page 63.)

[129] D. T. Westwick and R. E. Kearney. *Identification of nonlinear physiological systems*. IEEE Press series on biomedical engineering. IEEE Press, 2003. ISBN 0471274569 (cloth). URL http://www.loc.gov/catdir/bios/wiley044/2003043255.html. (Cited on page 50.)

[130] Z. Wu, N. E. Huang, S. R. Long, and C.-K. Peng. On the trend, detrending, and variability of nonlinear and nonstationary time series. *Proceedings of the National Academy of Sciences*, 104(38): 14889–14894, 2007. (Cited on page 28.)

[131] Z. Zi, Y. Zheng, A. Rundell, and E. Klipp. SBML-SAT: a systems biology markup language (SBML) based sensitivity analysis tool. *BMC Bioinformatics*, 9(1):342, 2008. ISSN 1471-2105. doi: 10.1186/1471-2105-9-342. URL `http://www.biomedcentral.com/1471-2105/9/342`. (Cited on pages 101 and 105.)